Fall 12-18-2020

Determination and Dissection of DNA-binding Specificity for the Thermus thermophilus HB8 Transcriptional Regulator TTHB099

Kristi Moncja

Determination and Dissection of DNA-binding Specificity for the *Thermus thermophilus* HB8 Transcriptional Regulator TTHB099

by

Kristi Moncja

BS in Biochemistry

Kennesaw State University, 2019

Submitted in Partial Fulfillment of the Requirements

For the Degree of Master of Science in the

Department of Chemistry and Biochemistry

Kennesaw State University

2020

DocuSigned by:
Michael Van Dyke
FFF164E8A9194D8...

Committee Chair

DocuSigned by:
Chris Dockery
C1295244895A414...

Graduate Program Coordinator

DocuSigned by:
Melanie Griffin
37F843D38AFE42D...

Committee Member

DocuSigned by:
Mark Mitchell
88C4A21FAB064E5...

Department Chair

DocuSigned by:
Glen Meades
85C95E66610D4F5...

Committee Member

DocuSigned by:
Kojo Mensa-Wilmot
39226A8692FB45B...

College Dean

ACKNOWLEDGMENTS

I wish to express my deepest gratitude to my supervisor, Dr. Michael Van Dyke, for encouraging and guiding me through my graduate studies. I could not have chosen a better mentor! I wish to thank Dr. Glen Meades and Dr. Melanie Griffin for serving as members of my research committee, for their guidance and valuable input to better this project. Thank you for your support to the faculty and staff in the Department of Chemistry and Biochemistry, the College of Science and Mathematics, and the Graduate College at Kennesaw State University. Thank you to the US National Science Foundation for financially supporting this project (Grant MCB 1714778).

I would like to thank all my friends and the Van Dyke group for their encouragement and scientific input. Moreover, I would like to thank my families for teaching me the value of education, inspiring and helping me through this journey. Special thanks to my fiancé, Brett Patterson – your love for knowledge is contagious and inspiring. Finally, a very big thank you to everyone who will read this document in their glorious quest for scientific knowledge.

ABSTRACT

Transcription factors (TFs) have been extensively researched in certain well-studied organisms but far less so in others. Following the whole-genome sequencing of a new organism, TFs are typically identified through their homology with related proteins in other organisms. However, recent findings demonstrate that structurally similar TFs from distantly related bacteria are not usually evolutionary orthologs. Here we explore TTHB099, a cAMP receptor protein (CRP)-family TF from the extremophile *Thermus thermophilus* HB8. Using the *in vitro* iterative selection method Restriction Endonuclease Protection, Selection and Amplification (REPSA), we identified the preferred DNA-binding motif for TTHB099, 5'-TGT $\left(\frac{A}{g}\right)$ n $\left(\frac{t}{c}\right)$ c $\left(\frac{t}{c}\right)$ $\left(\frac{a}{g}\right)$ g $\left(\frac{a}{g}\right)$ n $\left(\frac{T}{c}\right)$ ACA-3', and mapped potential binding sites, and regulated genes within the *T. thermophilus* HB8 genome. Comparisons with expression profile data in TTHB099-deficient and wild type strains suggested that, unlike *E. coli* CRP (CRP_{Ec}), TTHB099 does not have a simple regulatory mechanism. However, we hypothesize that TTHB099 can be a dual-regulator similar to CRP_{Ec}.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

BLI	Biolayer Interferometry
DBD	DNA-binding Domain
DOOR^2	Database of Prokaryotic Operons
cAMP	Cyclic Adenosine Monophosphate
CRP	Cyclic-AMP Receptor Protein
EMSA	Electrophoretic Mobility Shift Assay
ETC	Electron Transport Chain
FIMO	Find Individual Motif Occurrences
FNR	Fumarate and Nitrate Reductase Regulatory Protein
GEO	Gene Expression Omnibus
HTH	Helix-Turn-Helix
IISRE	Type IIS Restriction Endonuclease
ISP	Individual Sequencing Particle
IPTG	IsoPropyl β-D-1-thiogalactopyranoside
KEGG	Kyoto Encyclopedia of Genes and Genomes
LB	Lysogeny Broth
MEME	Multiple Em for Motif Elicitation
NAPs	Nucleotide-Associated Proteins

NCBI	National Center for Biotechnology Information
PAGE	Polyacrylamide Gel Electrophoresis
PCR	Polymerase Chain Reaction
PGM	Personal Genome Machine
REPA	Restriction Endonuclease Protection Assay
REPSA	Restriction Endonuclease Protection, Selection, and Amplification
RNAP	Ribonucleic Acid Polymerase
TF	Transcription Factor
TFBS	Transcription Factor Binding Site
TSS	Transcription Start Site

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

1.1 The Apparatus of Bacterial Transcription

Groundbreaking genome sequencing projects over the past four decades accompanied by *in vivo, in vitro,* and *in silico* studies have led to new understandings of biological processes in many prokaryotic model organisms such as *Escherichia coli* (*E. coli*), *Bacillus subtilis (B. subtilis*), *Pseudomonas aeruginosa* (*P. aeruginosa*), and *Thermus thermophilus* (*T. thermophilus*). Throughout all prokaryotic life, one of the main biological processes that control growth, proliferation, and adaptive responses is the regulation of gene expression. Gene expression starts with the transcription of DNA into RNA, the process in which an RNA polymerase (RNAP) complex binds to a unique DNA sequence known as a promoter, and proceeds to create an RNA copy of the DNA segment being transcribed. In bacteria, such a process is regulated by different factors: (a) topology of promoters and their recognition by RNAP, (b) concentration of free active RNAP, and (c) the presence of transcription factors (TFs) and their small molecule modulators [1].

Promoters are DNA sequences located upstream of the transcription starting site (TSS), where the RNAP complex binds to control gene expression. In *E. coli*, the two principal promoters are the –35 and –10 motifs, the TTGACA and TATAAT hexamers located approximately 35 and 10 bp upstream of the TSS [2]. Additional RNAP interactions are mediated through the upstream (UP) elements made of adenine (A), and

thymine (T) repeats, the extended –10 elements, as well as the spacer elements [2–4]. These promoter elements were identified to interact with E. coli's main sigma factor, sigma 70 (σ^{70}). However, bacteria have alternative σ factors, hence about 2,000 variations of promoters [2]. Such disparity reveals one degree of transcription control. Promoter similarity to its consensus sequence has been used to recognize promoter strength, indicating higher transcription rates. Nevertheless, the transcription rate depends on other factors apart from promoter homology to its consensus sequence. One factor is DNA topology, such as DNA supercoiling or nucleotide-associated proteins (NAPs) [5]. Other factors are levels of RNAP and TFs.

Bacterial RNAP holoenzymes are comprised of the core enzyme ($\alpha_2\beta\beta'\omega$) and one σ factor. The core enzyme has two identical 329-residue alpha subunits (α_2), with each subunit having two independently folded domains (larger alpha amino-terminal domain (α-NTD) and smaller alpha carboxy-terminal domain (α-CTD)). Additionally, the core contains the large beta (β) and beta prime (β′) subunits (1,342 and 1,407 residues, respectively) as well as a small 91 residue omega (ω) subunit [2]. Once the holoenzyme is formed, it can recognize promoter regions, interact with TFs, and start transcription. After synthesizing about 9 to 12 nucleotides of RNA, other interactions between RNAP and DNA allow for elongation initiation. The last step, termination, results in the RNAP separating from DNA and the newly synthesized RNA, followed by the core enzyme dissociating from the σ factor [6]. Units of RNAP in the cell are in short supply, and sometimes the active enzyme is bound to DNA non-productively. Moreover, different σ factors compete with one another to form the RNAP holoenzyme. Therefore, different

promoters compete for RNAP holoenzymes [2]. To further increase the control on which genes are transcribed, bacteria use TFs to activate or repress transcription.

Transcription factors are *trans* factors that bind to *cis*-regulatory elements as well as other *trans* factors. It has been reported that most of the bacterial *cis*-regulatory elements are found in the proximal region (about –100 to +20 bp from TSS) and distal regions or enhancers (up to –200 bp from TSS) [7–9]. Structurally, the majority of bacterial TFs have two domains: the DNA-binding domain (DBD) and the regulatory domain (RD), also known as the companion domain (CD). The role of the secondary domain is to interact with RNAP and other TFs [10,11].

TFs establish sequence-specific DNA interactions through their DBD. This stretch of amino acid (aa) residues determines a TFs' interactions with a specific DNA sequence known as the transcription factor binding site (TFBS). For example, many transcription factors bind their TFBSs with nanomolar affinity, while others exhibit micromolar attractions. Moreover, some TFs regulate transcription by promoting a configurational change of the DNA like a 90-degree kink. Based on the DNA-binding domains' structural and functional characteristics, genome comparison studies have categorized most bacterial DBDs to belong to the helix-turn-helix (HTH) family [12–14]. This roughly 60-64 aa domain primarily interacts with DNA's major groove via the secondary α-helix [15]. Most HTH transcription factors recognize palindromic DNA regions [16]. In many cases, TFs tend to dimerize, trimerize, or tetramerize to increase binding specificity to DNA.

Via their regulatory domains, interactions between TFs and RNAP establish another level of control over transcription. The regulatory domains interact with the RNAP's α-NTD or with its α-CTD and one of the σ factors that recruit RNAP to specific promoters [2]. These interactions are sometimes dependent on the location and orientation of the TFBS and result in guiding RNAP to the promoter or helping it bind tighter to DNA [17]. In other transcription factors, known as secondary channel-binding factors (SCBFs), regulatory domains interact with the β′ subunit of the RNAP holoenzyme. Although SCBFs are not essential to cells' natural growth, and they are absent in some bacterial genera, their interactions with RNAP seem to prevent conflicts between the replication fork and elongating RNAP [18].

Having established the interactions of TFs with both DNA and RNAP, which eventually make up the response to stimuli, it is time to explore how TFs recognize intracellular and extracellular changes such as nutritional (biomolecules, ions, minerals) and physiochemical (light, temperature, pressure, redox potential, oxygen content, water content) [19]. TFs utilize their regulatory domains, also known as signal-sensing domains (SSDs), to bind unique small-molecule modulators such as cyclic nucleotide monophosphate (c-NMP), flavin adenine dinucleotide (FAD), hydrogen peroxide (H_2O_2), nitric oxide (NO), carbohydrate metabolites, derivatives of enzymes, and metals [20–24]. Interactions of TFs with their signaling ligands usually induce conformational changes within TFs that promote or inhibit their ability to bind DNA. In addition to DBD comparisons, RD similarities are used to further group TFs into families. For example, the HTH TFs can be further arranged into superfamilies such as TetR/AcrR, GntR, and CRP/FNR, to name a few [11]. Databases like Pfam, Superfamilies, and Prosite use

sequence, structural, and functional information to find homologous TFs through the tree of life [15]. In turn, these findings can be used for TF identification and gene regulation investigation via biochemical studies.

1.2 Types of Transcription Regulators

TFs are classified into global and local regulators based on the number and function of genes they monitor. Global TFs regulate a sizable number of operons that belong to different metabolic pathways. Global regulators in *E. coli,* seven families of TFs (CRP, FNR, IHF, FIS, ArcA, NarL, and Lrp), are responsible for regulating 51% of the genes [25]. Within this group, cyclic-AMP receptor protein (CRP) regulates a total of 197 genes, FNR controls 111 genes mostly involved in nitrogen metabolism, ArcA regulates 63 genes of aerobic respiratory control, and NarL controls 65 genes of nitrate/nitrite regulation and anaerobic respiration. In *B. subtilis,* six TFs were identified as global regulators controlling diverse cellular processes such as aa biosynthesis, energy, and transport [26]. The rest of the TFs in these two model organisms were recognized as local or specific TFs responsible for regulating genes from a single pathway or belonging to the same functional classification. Furthermore, many global regulators interact with local regulators, forming the mechanism of co-regulation. In many cases, a co-regulatory system produces a feedback loop influencing its expression. Distinguishing negative and positive feedback loops is achieved by recognizing which TFs are activators and which are repressors.

TFs that stimulate the transcription of the genes they govern are called activators. Four simple activation methods encompass most forms of transcription activation in bacteria [2,27,28]. The first, known as Class I activation, involves the binding of the TF upstream of the –35 promoter element and interacting with the α-CTD component of RNAP, recruiting the enzyme to the specific promoter [2]. The second, or Class II

mechanism, consists of the TF binding on the –35 element and interacting with the σ factor for recruiting RNAP to the specific promoter. The third simple activation mechanism involves a TF binding near or on the promoter (either –35 or –10 box), inducing a conformational change of the DNA sequence and accommodating the promoter to be recognized by the σ factor of RNAP [2]. This third mechanism does not involve direct TF-RNAP interactions. The fourth mechanism is called activation via repressor modulation, where an activator binds a repressor, thus interfering with the repressed state of transcription. Lastly, there are reports of a few activators binding downstream of the –10 promoter element, but their regulatory mechanism is yet to be understood [29].

On the contrary, repressors are those TFs that reduce transcription. There are four distinct mechanisms used to describe transcription repression [2,27,28]. Steric hindrance is one of the most acceptable methods, in which a TF binding site is located on or between the core promoters (–35 and –10 box) and prevents RNAP from binding to the promoter. Repression by looping, the second method, does not prevent RNAP from binding to the promoter but instead induces looping of DNA, which shuts off transcription initiation. The third method, repression by modulation of activators, uses TF-TF interactions where repressors bind activators and prevent them from initiating transcription. The fourth method is called the roadblock mechanism, in which a TF binds at the start of the coding region and blocks transcription elongation. In a few cases, suppressors can bind upstream of promoters promoting RNAP holoenzyme docking via protein-to-protein interactions [30].

Furthermore, the same TFs can act as both activators and repressors, depending on where they bind regarding promoters and how they interact with RNAP. Most global TFs in bacteria are dual regulators. A simple dual regulation method is observed when the TFBS is located in the intergenic region of divergent operons. Such a theme is present in sugar metabolism loci, in which structural genes are activated while the TF gene is repressed. Another method is via the interplay between TF concentration and binding sites strength. A dual regulator can have a strong TFBS near a promoter and a weak TFBS inside the promoter. At low concentrations, the dual TF will bind to the strong TFBS and activate transcription. At high concentrations, the strong TFBS will be saturated, and excess TF will bind the weak TFBS, thus repressing transcription via steric hindrance [31]. The dual nature of some TFs allows for genetic resource conservation by using intricate regulatory mechanisms.

1.3 Significance of TF Identification

Understanding transcription regulation in bacteria provides insights into how these organisms adapt to various environmental niches and how they compare to one another. For instance, the number of TFs varies with genome size: the larger the genome, the more regulators are present per gene [32]. Point mutation studies combined with bioinformatic studies of evolutionary events, such as gene duplication and horizontal gene transfer, have also shed light on adaptation [33]. Analogous studies have shown that the numbers of TFs fluctuate with the organism's lifestyle. Free-living bacteria tend to have a higher number of TFs compared to the strict parasitic ones, and bacteria with complex life cycles, such as those with free-living and parasitic (e.g., *P. aeruginosa*) or symbiotic (e.g., *S. meliloti*) stages, tend to have a high proportion of TFs in general [13]. Furthermore, the conservation of TFs between bacterial species, but not between prokaryotes and eukaryotes, has prompted many novel antibacterial drug developments [34].

The main approach in understanding transcription regulation by TFs is by identifying the TFs' DNA binding motifs. Technological advances have resulted in the twining of bioinformatics and biochemistry into *in vivo* and *in vitro* contemporary combinatorial techniques. *In vivo* methods are advantageous for analyzing TF binding events at different time points or under specific conditions [35]. One method is the genetic manipulation of the TF of interest, followed by DNA microarray studies. However, there have been false positive and false negative results primarily when the TF of interest affects other TFs. Chromatin immunoprecipitation paired with high-throughput sequencing (ChIP-seq) is a widely used method in which the organism’s chromatin after

being exposed and chemically cross-linked to the TF of interest *in vivo*, is then fragmented and the TF-DNA fragments are immunoprecipitated via a TF specific antibody. Nevertheless, ChIP can lead to the selection of binding sites much larger than the TFBSs themselves, and these motifs need to be discovered within the selected sequences.

In vitro techniques are preferred for large-scale characterization of intrinsic TFBS and *de novo* motif discovery [35,36]. Protein-binding microarrays (PBMs) and High-throughput Systematic Evolution of Ligands by Exponential Enrichment (HT-SELEX) are *in vitro* methods that rely on TF selection of high-affinity binding sites from a large pool of libraries [37,38]. These techniques rely on the physical-based or affinity-based separation of the TF-DNA complexes.

Restriction Endonuclease Protection, Selection, and Amplification (REPSA) is a novel *in vitro* combinatorial method developed by the Van Dyke Laboratory that does not require any affinity-based separation [39]. REPSA is a PCR based technique utilizing the selection of high-affinity TF-DNA interactions in a pool of randomized sequences, extricating the unbound DNA sequences via type IIS restriction endonuclease (IISRE) activity, and amplifying the preferred sequences for further selection.

One of the key components of REPSA is its selection template, derived from the ST2R24 template precursor (Figure 1) [40,41]. It is composed of a 23-mer (ST2L) primer on the 5' end and a fluorescently red-labeled 25-mer (IRD7_ST2R) primer on the 3' end, flanking a 24-mer randomized region. The randomized region was designed to contain about 42 femtomoles or 2.5×10^{10} different DNA molecules, large enough in both size

and variation, to provide a preferred binding sequence for many prokaryotic transcription factors. The IRDye® 700 (IRD7)-label on the right primer allows for visualization of results after polyacrylamide gel electrophoresis (PAGE). The IRD7_ST2R primer was designed to contain the two binding sites for the next important element of REPSA, type IIS restriction endonucleases: here FokI (CATCC) and BpmI (CTCCAG).

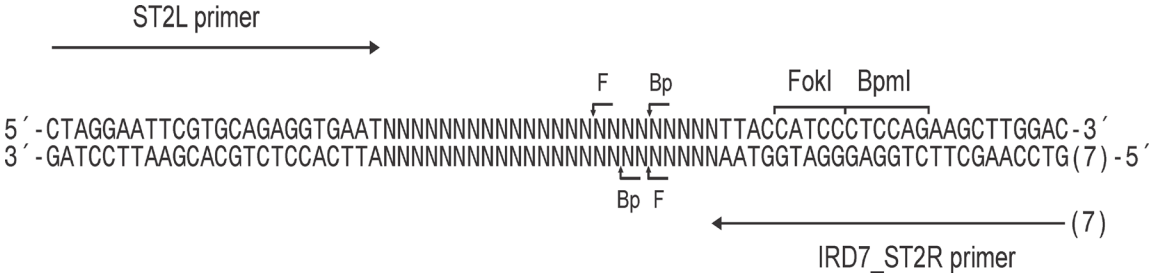

Figure 1. REPSA selection template. An illustration of the components of the REPSA template; horizontal arrows indicate ST2L primer and IRD7_ST2R primer; (N) represents random nucleotides within the 24 bp randomized region; brackets and arrows show the IISREs (FokI and BpmI) binding and cleavage sites, respectively. Adapted from [40].

Once the IISRE binds the template on the defined IRD7-ST2R primer, it would cleave the DNA in the randomized region at a specific distance without any cleavage-site specificity, as indicated by arrows in Figure 1. However, in the presence of a DNA-binding molecule, the transcription factor would bind the most preferred variation of the 24-mer randomized region, blocking the IISRE from cleaving that percentage of the DNA pool (Figure 2). The PCR rounds would then only amplify the protected sequences. The next rounds of REPSA would be seeded with sequences from the protected pool of the previous round, further refining the preferred binding site.

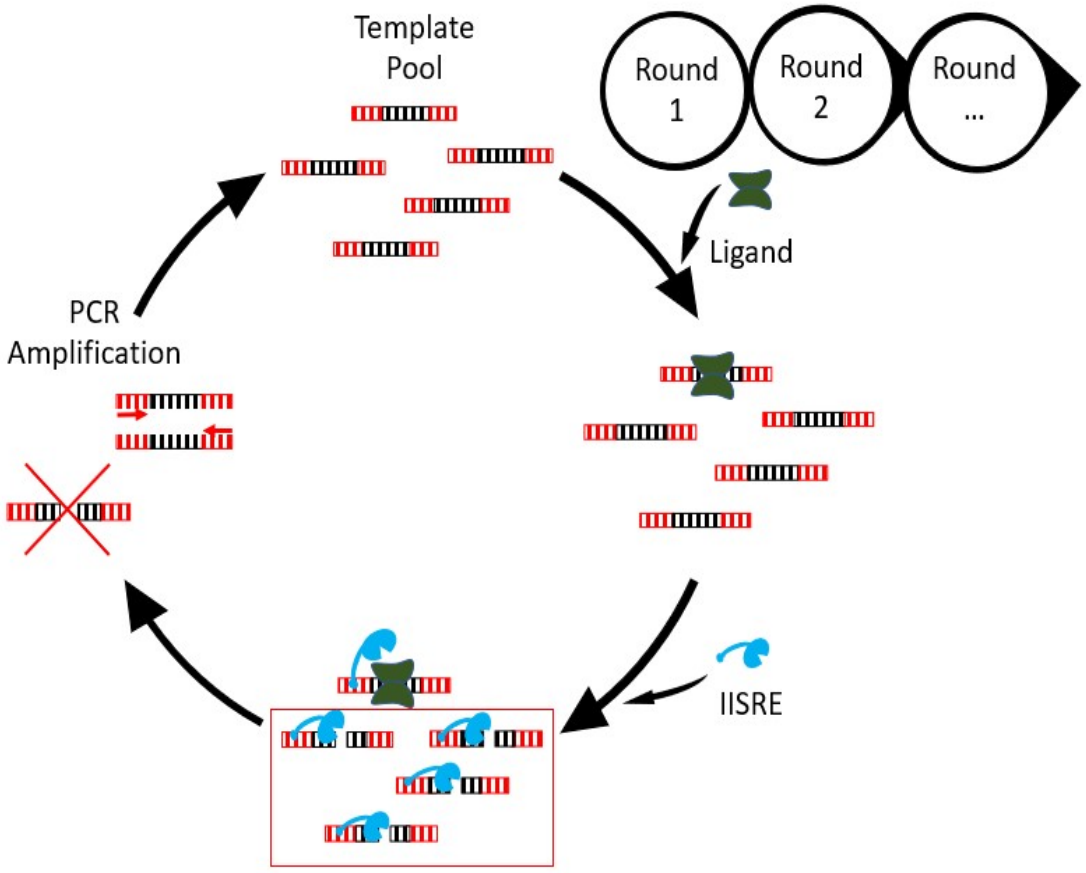

Figure 2. Diagram depiction of the REPSA assay. A step by step representation of the REPSA method, including its components: DNA pool symbolized by the red (primers) and black (randomized 24-mer) sequences; ligand is shown as a dimer in green; IISRE is shown in blue. **Step 1.** The introduction of the ligand in the template pool allows specific binding to a small percentage of DNAs. **Step 2.** The addition of IISRE in the previous reaction cleaves all the DNA sequences that are not bound to the ligand. **Step 3.** The reaction undergoes PCR, but only the uncut sequences are amplified. **Step 4.** The amplified sequences are used as a template for the next round of REPSA. Multiple rounds of REPSA will result in the ligand-specific selection of DNA sequences.

1.4 TF Discovery in *T. thermophilus* HB8

The Van Dyke laboratory has utilized REPSA to identify TFs from the extremophilic model organism *T. thermophilus* HB8. This gram-negative bacterium belongs to the Deinocooccus-Thermus phylum and grows in temperatures as low as 47 °C and as high as 85 °C, with an optimal range of 65-72 °C [42]. Its genome consists of a 1.85 megabase pair circular chromosome (TTHA), a 257 kilobase pair megaplasmid (TTHB or pTT27), and a 9.32 kilobase pair miniplasmid (TTHC or pTT8) [43]. This model organism has been the epicenter for the Structural-Biological Whole Cell Project at RIKEN Harima Institute in Japan. Studies in metabolic pathways and enzymes from *T. thermophilus* HB8 have been of significant importance for systems biology and industrial processes [44,45,46].

Our present study focuses on the characterization of TTHB099 TF. This is a putative dual functioning transcription regulator for which a DNA-binding motif has not been identified. Furthermore, there is a lack of understanding about its regulatory mechanism and biological role in *T. thermophilus* HB8. The TTHB099 gene is located within the TTHB megaplasmid, second in the *TTHB100* operon, also known as the litR operon [47,48]. The TTHB099 monomer is made of 195 aa, has a molecular mass of 22,138 Da, and an HTH motif (aa 142–161) [49]. It has been recognized as one of the four CRP/FNR superfamily members (TTHB099, TTHA1359, TTHA1437, and TTHA1567) in *T. thermophilus* HB8 [50]. The lack of cysteine residues indicates that TTHB099 cannot detect oxygen or redox variations by interacting with iron-sulfur clusters. Furthermore, despite having an effector domain, TTHB099 does not require cAMP to bind DNA [50]. Indeed, the crystal structure of TTHB099 without cAMP resembles that of *E. coli* CRP

(CRP_{Ec}) when bound to cAMP (Figure 3). There have been no reports on other small effector molecules that could potentially interact with TTHB099, leaving no explanation of how this TF detects cellular changes.

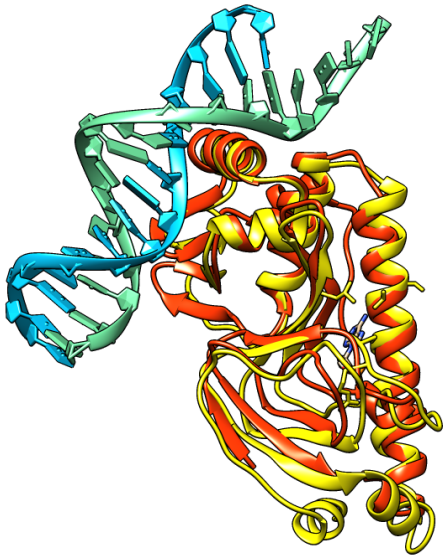

Figure 3. Superimposed crystal structures of TTHB099 and CRP_{Ec}. Image depicting superimposed crystal structures of CRP_{Ec} (yellow) bound to a CAP-DNA complex (blue/green) (PDB ID: 1J59 chain A); TTHB099 (orange, PDB ID: 3B02 chain A); analyzed in Chimera 1.14 using publicly available data from PDB [50–52].

Investigations into the biological role of TTHB099 have led to TTHB099's closest homolog, TT_P0055, from *T. thermophilus* HB27 (one aa substitution, E77D), and further analysis has shown that the respective genes and their upstream regulatory regions have 99% similarity [48]. Indeed *T. thermophilus* HB8 and HB27 are two closely related strains from the Thermus genus. Genome comparisons of the two strains (HB8 and HB27) revealed that they both contain a highly conserved chromosome region with 94% of the genes shared and an average of 97.6% aa identity [53]. One main difference between the two strains is that the *T. thermophilus* HB8 can grow anaerobically in the presence of nitrate, which is attributed to the additional mini plasmid (pTT8) [54].

TT_P0055 has been reported to be a positive regulator of the *crtB* operon, which in turn is involved in light-dependent carotenoid biosynthesis [48]. An 18-mer potential binding sequence for TT_P0055, AGTGT[N7]GCAAAA, was identified upstream of *crtB* operon. The study only focused on one location; hence this sequence does not represent the general TT_P0055-prefered DNA binding site in *T.thermophilus* HB27. Nevertheless, it is the only TFBS predicted for TT_P0055. Furthermore, due to this limited study revolving around only one operon, one cannot characterize TT_P0055 as a local or global regulator. There have been no reports of a potential TTHB099 binding sequence in *T. thermophilus* HB8. In this study, a reverse genetic approach will be used to ascertain TTHB099's TFBSs. Binding kinetics studies and bioinformatic analysis will be used to infer the regulatory mechanism and the biological role of TTHB099 in *T. thermophilus* HB8.

1.5 Hypothesis and Specific Aims

A reverse genetic approach can be used to ascertain the biological functions of transcriptional regulator TTHB099.

1. Express and purify the protein of interest, TTHB099.
2. Obtain TTHB099's consensus sequence using our novel selection method REPSA.
3. Validate and sequence the selected consensus DNA-binding site.
4. Map the consensus sequence into the genome of *T. thermophilus* HB8 to indicate which genes are regulated by TTHB099.
5. Validate potentially regulated genes by biophysical means using BLI.
6. Determine the biological functions of TTHB099 using a bioinformatic approach.

MATERIALS AND METHODS

2.1 Modified Oligonucleotides

The nucleic acid precursors and primers used in this study were synthesized and purified by Integrated DNA Technologies (https://www.idtdna.com), and are displayed in Table 1. A pool of single-stranded selection template, ST2R24, was designed to have an average nucleotide composition of the randomized cassette of 25% A, 25% C, 25% G, and 25% T at each position. The ST2R24 template precursor was then transformed into the double-stranded DNA pool via PCR. This step was comprised of five 25 µL reactions, each containing 1 ng single-stranded ST2R24 template precursor, 1X Standard *Taq* Reaction Buffer (New England Biolabs, NEB), 560 nM ST2L primer, 560 nM IRD7_ST2R primer, 50 µM dNTPs, and 25 U *Taq* DNA Polymerase, that were PCR amplified for seven cycles and consequently combined. Cycling conditions involved (5 cycles, 95°C/30 s, 60°C/30 s, 68°C/1 min; 1 cycle, 95°C/30 s, 60°C/30 s, 68°C/1.5 min; and 1 cycle, 95°C/30 s, 60°C/30 s, 68°C/2 min). This treatment should increase the amount of DNAs with a fully annealed random cassette and increase the randomized region's diversity.

Table 1. Nucleic acid precursors and primers used in this study.

Name	Sequence	Length	Purif.	Use
ST2R24	CTAGGAATTCGTGCAGAGGTGAATNNNNNNNNNNNNNNNNNNNNNNNNTTACCATCCCTCCAGAAGCTTGGAC	73	PAGE	ST2R24 Template Precursor

ST2L	CTAGGAATTCGTGCAGAGGTGAAT	24	Desalt	PCR Left Primer
ST2Ls	CTAGGAATTCGTGCAGAGGTGA	22	Desalt	PCR Left Primer Short
ST2R	GTCCAAGCTTCTGGAGGGATGGTAA	25	Desalt	PCR Right Primer
IRD7_ST2R	/5IRD700/GTCCAAGCTTCTGGAGGGATGGTAA	25	HPLC	5'-IRDye700 PCR Primer
ABC01_ST2R	CCATCTCATCCCTGCGTGTCTCCGACTCAGCTGCAAGTTCGATGTCCAAGCTTCTGGAGGGATG	64	PAGE	Fusion PCR Primer
trP1_ST2L	CCTCTCTATGGGCAGTCGGTGATCTAGGAATTCGTGCAGAGGTGA	45	PAGE	Fusion PCR Primer
A_uni	CCATCTCATCCCTGCGTG	18	Desalt	PCR Primer
trP1_uni	CCTCTCTATGGGCAGTCGG	19	Desalt	PCR primer
IRD8_trP1_ST2L	/5IRD800/CCTCTCTATGGGCAGTCGGTGATCTAG	27	HPLC	5'-IRDye800-modified PCR Primer
Bio_ST2R	/5BiodT/GTCCAAGCTTCTGGAGGGATG	22	HPLC	5'-biotinylated PCR primer
REPSAis	CTAGGAATTCGTGCAGAGGTGAATCGTCATAGAATTCGTTACCATCCCTCCAGAAGCTTGGAC	63	PAGE	REPSAis control DNA precursor
ST2_099_wt	AGGAATTCGTGCAGAGGTGAATTGTATTCTAGAATACATTACCATCCCTCCAGAAGCTTG	65	Desalt	TTHB099 consensus probe
ST2_099_wt_m1	AGGAATTCGTGCAGAGGTGAATGGTATTCTAGAATACATTACCATCCCTCCAGAAGCTTG	65	Desalt	TTHB099 mutant 1 probe precursor
ST2_099_wt_m2	AGGAATTCGTGCAGAGGTGAATTTTATTCTAGAATACATTACCATCCCTCCAGAAGCTTG	65	Desalt	TTHB099 mutant 2 probe precursor
ST2_099_wt_m3	AGGAATTCGTGCAGAGGTGAATTGAATTCTAGAATACATTACCATCCCTCCAGAAGCTTG	65	Desalt	TTHB099 mutant 3 probe precursor
ST2_099_wt_m4	AGGAATTCGTGCAGAGGTGAATTGTCTTCTAGAATACATTACCATCCCTCCAGAAGCTTG	65	Desalt	TTHB099 mutant 4 probe precursor
ST2_099_wt_m5	AGGAATTCGTGCAGAGGTGAATTGTACTCTAGAATACATTACCATCCCTCCAGAAGCTTG	65	Desalt	TTHB099 mutant 5 probe precursor
ST2_099_wt_m6	AGGAATTCGTGCAGAGGTGAATTGTATACTAGAATACATTACCATCCCTCCAGAAGCTTG	65	Desalt	TTHB099 mutant 6 probe precursor
ST2_099_wt_m7	AGGAATTCGTGCAGAGGTGAATTGTATTTTAGAATACATTACCATCCCTCCAGAAGCTTG	65	Desalt	TTHB099 mutant 7 probe precursor
ST2_099_wt_m8	AGGAATTCGTGCAGAGGTGAATTGTATTCAAGAATACATTACCATCCCTCCAGAAGCTTG	65	Desalt	TTHB099 mutant 8 probe precursor
ST2_099_0080(0081)p	AGGAATTCGTGCAGAGGTGAATTGTGTTTTAGTTTACTTTACCATCCCTCCAGAAGCTTG	60	Desalt	*TTHA0080(0081)* promoter DNA probe precursor
ST2_099_0030p	AGGAATTCGTGCAGAGGTGAATTGTGTACGAAATTACATTACCATCCCTCCAGAAGCTTG	60	Desalt	*TTHA0030* promoter DNA probe precursor
ST2_099_0506(0507)p	AGGAATTCGTGCAGAGGTGAATTGTTTTTCAAGATACATTACCATCCCTCCAGAAGCTTG	60	Desalt	*TTHA0506(0507)* promoter DNA probe precursor
ST2_099_0132(0133)p	AGGAATTCGTGCAGAGGTGAATTGTAAGGGAGAATAAATTACCATCCCTCCAGAAGCTTG	60	Desalt	*TTHA0132(0133)* promoter DNA probe precursor
ST2_099_C002(C003)p	AGGAATTCGTGCAGAGGTGAATTGTGAGTTATCTCACTTTACCATCCCTCCAGAAGCTTG	60	Desalt	*TTHC002(C003)* promoter DNA probe precursor
ST2_099_B088(B089)p	AGGAATTCGTGCAGAGGTGAATGGTAGCCTGGACCACATTACCATCCCTCCAGAAGCTTG	60	Desalt	*TTHB088(B089)* promoter DNA probe precursor

ST2_099_0647p	AGGAATTCGTGCAGAGGTGAATGGTAGCCAGGGATACATTACCATCCCTCCAGAAGCTTG	60	Desalt	*TTHA0647* promoter DNA probe precursor
ST2_099_1833p	AGGAATTCGTGCAGAGGTGAATTGTAGGCCAGGCCACGTTACCATCCCTCCAGAAGCTTG	60	Desalt	*TTHA1833* promoter DNA probe precursor
ST2_099_0641p	AGGAATTCGTGCAGAGGTGAATCGTGTCCCTGAACACATTACCATCCCTCCAGAAGCTTG	60	Desalt	*TTHA0641* promoter DNA probe precursor
ST2_099_0645p	AGGAATTCGTGCAGAGGTGAATTGTGCCTTTGGCCACATTACCATCCCTCCAGAAGCTTG	60	Desalt	*TTHA0645* promoter DNA probe precursor
ST2_099_1911(1912)p	AGGAATTCGTGCAGAGGTGAATTGTACTTGAGCATACCTTACCATCCCTCCAGAAGCTTG	60	Desalt	*TTHA1911(1912)* promoter DNA probe precursor
ST2_099_B003(B004) p	AGGAATTCGTGCAGAGGTGAATTGTAGCCCAGGCCAAATTACCATCCCTCCAGAAGCTTG	60	Desalt	*TTHB003(B004)* promoter DNA probe precursor
ST2_099_0201(0202)p	AGGAATTCGTGCAGAGGTGAATTTTGTTATACGCCACATTACCATCCCTCCAGAAGCTTG	60	Desalt	*TTHA0201(0202)* promoter DNA probe precursor
ST2_099_0374p	AGGAATTCGTGCAGAGGTGAATAGTGATGTAAACTAAATTACCATCCCTCCAGAAGCTTG	60	Desalt	*TTHA0374* promoter DNA probe precursor
ST2_099_0326p	AGGAATTCGTGCAGAGGTGAATTGTGTTGCAGGACCCATTACCATCCCTCCAGAAGCTTG	60	Desalt	*TTHA0326* promoter DNA probe precursor
ST2_099_1626(1627)p	AGGAATTCGTGCAGAGGTGAATGGTATGGGAAGCTACATTACCATCCCTCCAGAAGCTTG	60	Desalt	*TTHA1626(1627)* promoter DNA probe precursor

The resulting DNAs were run in a 10% Native PAGE (1X Tris-borate-EDTA, pH 8.3 at 25°C, 9:1 acryl:bis) for 10 min at 50V, then 1h at 102V, and imaged by LI-COR Odyssey Imager. Their concentration was measured with Qubit 3 Fluorometer following our published protocol [55].

Libraries for massively parallel semiconductor sequencing were prepared by a two-step fusion PCR process, using fusion primers A_BC01_ST2R and trP1_ST2L as the initial set and A_uni and trP1_uni as the second set (Table 1). A 25 µL reaction containing 2 µL DNA from REPSA Round 7, 1X Standard *Taq* Buffer (NEB), 50 µM dNTPs, 200 nM trP1_ST2L primer, 200 nM A_BC01_ST2R primer, and 1 U *Taq* DNA Polymerase (NEB) was PCR amplified for seven cycles (95°C/30 s, 54°C 30/s, and 68°C/1 min). Three 25 µL reactions identical to the one used for fusion PCR were seeded

with 2 µL resulting DNA and PCR amplified for 6, 9, and 12 cycles under the same cycling conditions as the previous experiment. Treated libraries were run in 10% Native PAGE for 10 min at 50V, then 1h at 102V, and stained with 2.5 µg/µL ethidium bromide for 10 min then destained for another 10 min in water. The gel was imaged via ultraviolet (UV) exposure using a Gel Doc™ EZ (BIO-RAD) instrument.

Electrophoretic mobility shift assay (EMSA) probes were PCR amplified with 220 nM ST2L primer and 180 nM IRD7_ST2R primer in 50 µL reactions containing 1 µL DNA template precursor, 1X Standard *Taq* Reaction Buffer, 50 µM dNTPs, and 2 U *Taq* DNA Polymerase, for 30 cycles (95°C/30 s, 58°C/30 s, 68°C/1 min). Similarly, nucleic acids used in BLI assays were amplified and, at the same time, biotinylated with 220 nM ST2L primer and 180 nM Bio_ST2R primer.

The control restriction endonuclease protection assay (REPA) probe was generated via a two-step method. The first step included a 25 µL reaction containing 2 ng REPSAis template precursor, 1X Standard *Taq* Reaction Buffer (New England Biolabs, NEB), 200 nM trP1_ST2L primer, 200 nM ST2R primer, 200 µM dNTPs, and 2.25 U *Taq* DNA Polymerase. This reaction underwent six PCR cycles under the following conditions (95°C/30 s, 58°C/30 s, 68°C/1 min). A second 25 µL PCR reaction (1X Standard *Taq* Reaction Buffer, 200 nM IRD8_trp1_ST2L primer, 200 nM ST2R primer, 200 µM dNTPs, and 0.625 units *Taq* DNA polymerase) was seeded with 1µL template from the previous reaction. This reaction underwent 30 cycles of PCR (95°C/30 s, 60°C/30 s, 72°C/1 min). The resulting DNAs were run in 10% Native PAGE, as previously described. The concentrations for each modified oligonucleotide were measured with Qubit 3 Fluorometer, as indicated above.

2.2 TTHB099 Protein Preparation

E. coli BL21 (DE3) competent cells were transformed with the PC014099-42 plasmid containing the *TTHB099* gene (obtained from RIKEN Bioresource Research Center) in Lysogeny broth (LB) media in the presence of 100 μg/ml ampicillin. The culture was incubated at 37°C, induced with 1 mM final concentration of isopropyl β-D-1-thiogalactopyranoside (IPTG), pelleted by centrifugation (4,000 rpm for 15 min/4°C), and resuspended in 0.5 mL 2X BEB (40 mM Tris-Cl [pH 7.5], 200 mM NaCl, 0.2 mM EDTA, 2 mM DTT, 1 mM PMSF). The cells were then lysed by three cycles of sonication (3 W/cm^2, 10 s on/10 s off, 0°C) and centrifuged (13,000 rpm for 10 min/RT). Further purification was done simply by heat. The lysed cells were heat-treated (70°C) for 15 min. Under such conditions, the *E. coli* proteins were denatured and were no longer soluble; however, the thermophilic TTHB099 protein was not affected. The heated sample was centrifuged (4,000 rpm for 15 min/4°C). The supernatant was retrieved, diluted with an equal volume of glycerol, and stored at –20°C. Protein purity and quantification were determined by 12% SDS-PAGE. Quantitative densitometry following staining with Coomassie Brilliant Blue G250 dye was done using a BSA standard curve (0.5, 1.0, and 2.0 mg/mL).

2.3 REPSA and Sequencing

REPSA 20 µL selections were performed with 4.515 ng (100 fmol) template DNA pool in 1X CutSmart® Buffer NEB (50 mM Potassium Acetate, 20 mM Tris-acetate, 10 mM Magnesium Acetate, 100 µg/ml BSA, pH 7.9 at 25°C). The first REPSA round was incubated with the double-stranded ST2R24 library, while the subsequent rounds were seeded with 2 µL DNA from the previously selected template. The DNA and IISRE cleavage controls (–/– and –/F or –/B, respectively) contained 1 µL PDB (20 mM Tris HCl, 100 mM NaCl, 0.1 mM EDTA, 1 mM DTT, 100 µg/mL BSA, 0.1 % Tween 20, pH 8.0 at 25°C) instead of the TTHB099 ligand, while the experimental (+/F) was incubated with 50.6 nM purified TTHB099 protein. These reactions were incubated for 20 min at 55°C and equilibrated for 5 min at 37°C. The DNA control was treated with 0.8 µL PDB, while the IISRE control and the experimental reactions were treated with 3.2 U FokI and 8 U BpmI enzymes for Rounds 1–4 and 5–7, respectively. The reactions were incubated for 5 min at 37°C to allow for DNA cleaving and placed on dry ice for 2 min to terminate the endonuclease activity.

The REPSA amplification step involved three 23 µL reactions containing 1X NEB standard *Taq* Reaction Buffer, 50 µM dNTPs, 200 nM primers ST2L, 200 nM IRD7_ST2R, and 5 U NEB *Taq* DNA polymerase assembled on ice. The three 23 µL reactions incubated with 2 µL from the selection reaction were then PCR amplified for 6, 9, and 12 cycles under the following protocol: 30 s denaturation at 95°C, 30 s annealing at 58°C, and 60 s elongation at 68°C. Following PCR amplification, 2 µL aliquots from amplified reactions were mixed with 2 µL 6X Orange Gel Loading Dye (20% wt/vol dextrose, 0.9% wt/vol Orange G dye, 1% wt/vo SDS, and 66 mM EDTA). A 10% Native

PAGE was run in 5X TBE for 10 min at 50V, then 1h at 102V, and imaged by LI-COR Odyssey Imager. DNA concentrations were measured by Qubit 3 Fluorometer following our published protocol.

The selected sequences obtained from REPSA were validated using REPA. This method was run very much like the selection step of REPSA [56], with the addition of a green fluorescently labeled control DNA (REPSAis). REPSA results were massively parallel sequenced using Thermo Fisher Ion Personal Genome Machine (Ion PGM), as previously described [40]. The sequencing results were selected by the Sequencing1.java program to contain only the sequences with intact flanking primers and a randomized region of 24 bp in length. A set of 1,000 reads from the refined data were further analyzed by the web version 5.0.5 of Multiple Em for Motif Elicitation (MEME) (http://meme-suite.org/tools/meme) [57], using default parameters with and without a palindromic filter. The MEME results, position weight matrixes displayed as sequence logos, helped identify the 16-mer preferred consensus sequence selected by REPSA.

2.4 Binding Assays

EMSA 10 μL reactions were performed with DNA libraries from REPSA Round 1 and Round 7 selections [58]. Each reaction contained 1X Cutsmart Buffer (NEB), 2 ng DNA, as well as 2 μL TTHB099 protein corresponding to the following ten-fold serial dilutions (0, 5.06, 50.6, 506, and 5,060 nM). All ten reactions were incubated at 55°C for 20 min to promote binding, then at 37°C for 5 min to stabilize the DNA-protein complex. The 2 μL samples were mixed with 2 μL 6X Orange Gel Loading Dye without SDS (20% wt/vol dextrose, 0.9% wt/vol Orange G dye, and 66 mM EDTA) and loaded in a 0.5X TAE, 10% wt/vol polyacrylamide (19:1 acryl:bis) gel. The gel was run in 5X TAE for 10 min at 50V, then 1h at 102V. Results were visualized by IR fluorescence as previously described.

A second EMSA was run similarly to the first one to test the binding of TTHB099 to its defined consensus sequence. This time, the 10 μL reactions were seeded with 1.1 nM ST2_099 DNAs and 2 μL from two-fold serial dilutions of TTHB099 protein (0, 0, 0.66, 1.32, 2.64, 5.27, 10.5, 21.1, or 42.2 nM). The results were visualized and quantified using a LI-COR Odyssey Imager.

Biolayer interferometry (BLI) was used to measure real-time binding kinetics for TTHB099 and various DNA probes. The assays were run in the FortéBio Octet^{QK} instrument in 96-well microplates using streptavidin sensors and biotinylated oligos. Each assay was designed with four lanes by four rows, containing 200 μL reactions buffered with BLI 100 buffer (20 mM Tris-HCl, 100 mM NaCl, 1 mM EDTA, 0.05% Tween 20, pH 7.7 at 25°C). The first lane contained 2 nM biotinylated DNAs and served as the loading step. The second and fourth lanes contained 200 μL BLI 100 buffer and

served as the background and dissociation steps, respectively. The third lane included four concentrations of TTHB099 (17, 50, 150, 450 nM), which provided the association step. The results from BLI were transferred in GraphPad Prism 8, where least squares regression analysis of the association and dissociation steps were used to derive binding parameters and graphs.

2.5 Bioinformatic Studies

The 16-mer position weight matrix data derived from MEME were inputted in Find Individual Motif Occurrences (FIMO)(http://meme-suite.org/tools/fimo) to map the identified sequences into the genome of *T. thermophilus* HB8 (GenBank uid13202 210) [59]. Only the results with P-values $\leq 3.95 \times 10^{-5}$ were further analyzed, similar to previous studies. Sequences ±200 bp from the 16-mer binding site were selected via the Kyoto Encyclopedia of Genes and Genomes (KEGG) (https://www.kegg.jp/kegg /kegg2.html) and examined for core promoter elements in Softberry BPROM (http://www.softberry.com/berry.phtml?topic=bprom&group=programs&subgroup=gfind b) [47,60]. Furthermore, operons were identified using the ProOpDB at the Universidad Nacional Autónoma de México (http://biocomputo.ibt.unam.mx: 8080/OperonPredictor/) and BioCyc (http://biocyc.org) [61,62]. Publicly available microarray data for gene expression profiles in wild-type and TTHB099-deficient *T. thermophilus* HB8 were obtained from the NCBI GEO website (https://www.ncbi.nlm.nih.gov/geo/) [63] (SuperSeries GSE21875). In particular, samples GSM532194, 5, 6, obtained from wild-type *T. thermophilus* HB8 grown in a rich medium for 360 min, and samples GSM530118, 20, 22, obtained from TTHB099-deficient *T. thermophilus* HB8 strains propagated under identical conditions. These data sets were analyzed using the NCBI GEO2R program with default settings to determine changes in gene expression (LogFC values) and their statistical significance (P-values).

RESULTS

3.1 TTHB099 Protein Expression, Purification, and Quantification

TTHB099 protein expressed in *E. coli* cells and purified via heat treatment was qualitatively and quantitatively assayed by a 12% SDS-PAGE. Fractions of bacterial proteins from each step of TTHB099 expression and purification are shown in Figure 4A. Following IPTG induction, TTHB099 can be visualized as a strong band with a molecular weight of about 22 kDa, consistent with the literature [50]. Further comparison of the band in the purified phase with the one in the soluble phase estimated that TTHB099 was greater than 90% pure (Figure 4A, lane 4). The presence of a few denatured *E. coli* proteins at low concentrations seen in lane 4 should not affect the later experiments in this study, as previously found for other *T. thermophilus* HB8 TFs studied in our laboratory [40,64–66]. The purified TTHB099 preparation had a concentration of 50.6 μM.

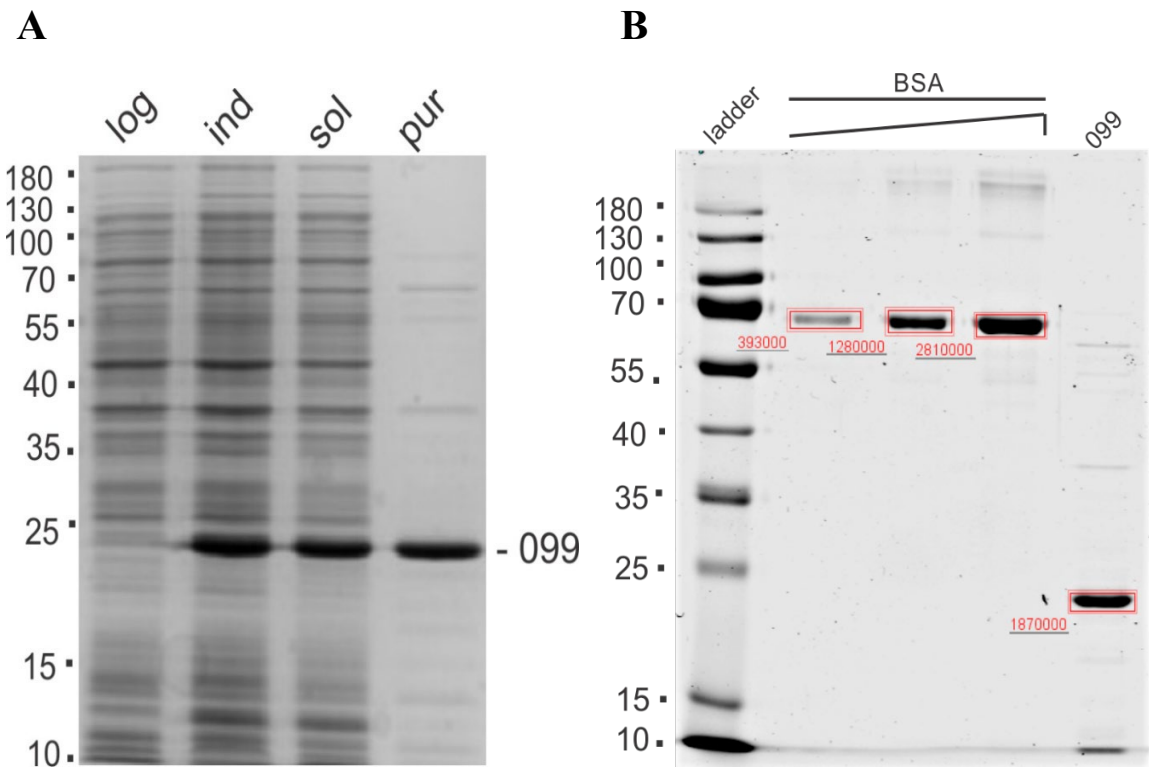

Figure 4. Expression, purification, and quantification of TTHB099 protein. (**A**) Shown is a Coomassie Blue G-250 stained 12% SDS-PAGE gel onto which was loaded whole-cell extracts or partially purified fractions equivalent to 0.2% of the total preparation. Lanes shown left to right: (log) logarithmic growth bacteria, (ind) bacteria following IPTG-induction for 4 h, (sol) soluble proteins following sonication and centrifugation, (pur) 2.3 µg purified TTHB099 protein. The location of molecular weight standards is indicated at the left of the figure. (**B**) Quantitative densitometric analysis of a Coomassie Blue G-250 stained 12% SDS-PAGE gel containing a BSA standard curve (left to right: 0.5, 1, 2 mg protein) and 0.5 µL stock TTHB099. The final concentration of TTHB099 is estimated to be 50.6 µM.

3.2 Determination of the TTHB099 DNA-binding Motif

REPSA was used to select the TTHB099 binding sites from a large pool of about 60 billion template molecules. This ST2R24 selection library has been successfully used in four previous TF identification studies [40,64–66]. Here, seven rounds of REPSA resulted in the emergence of DNA resistant to IISRE cleavage when TTHB099 was present (Figure 5, Round 7). For that round: the template in the DNA control (–/–) was uncut in the absence of BpmI and TTHB099; the template in the cleavage control (–/B) was cut entirely in the absence of TTHB099; the template in the experimental lane (+/B) was ~60% uncut in the presence of BpmI and TTHB099, representing the selected sequences. Note that the initial rounds of REPSA (1–4) were cleaved by FokI type IISRE. The emergence of the uncut template in the cleavage control (–/F) for Round 4 was attributed to the development of FokI cleavage-resistant DNAs. These cleavage-resistant DNAs had FokI binding motifs emerge in the randomized region that would interfere with proper FokI cleaving. In response, the following REPSA rounds (5–7) were cleaved by BpmI. The technique was modified this way to make use of the FokI's higher efficiency compared to BpmI. REPSA results were validated by REPA and EMSA, then sequenced.

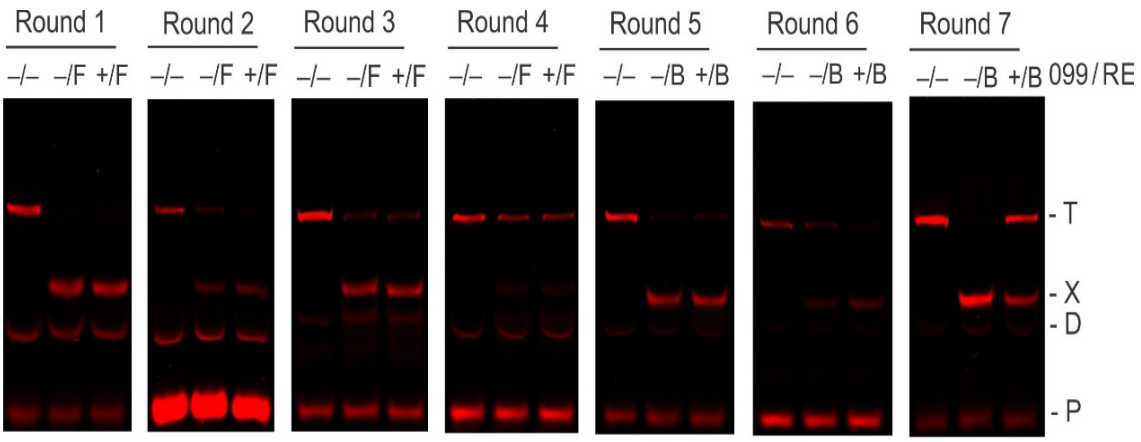

Figure 5. REPSA selection of TTHB099-binding DNA sequences. Shown are IR fluorescence images of restriction endonuclease cleavage-protection assays made during Rounds 1–7 of REPSA selection with 50.6 nM TTHB099 protein. The presence (+) or absence (–) of TTHB099 and IISRE FokI (F) or BpmI (B) are indicated above each lane. The electrophoretic mobility of the intact (T) and cleaved (X) ST2R24 selection template, primer dimer species (D), as well as the IRD7_ST2R primer (P) are indicated at the right of the figure.

REPA was performed for REPSA Rounds 4 and 7 to confirm the discrimination of TTHB099-specific and nonspecific IISRE cleavage inhibition (Figure 6A). A fluorescent green-labeled probe containing a defined template to which TTHB099 does not bind with high affinity, REPSAis, was introduced to the IISRE cleavage in the presence and absence of TTHB099. For both runs, the green control was cleaved by IISREs in a TTHB099-independent manner. The red-labeled test DNA followed the same trends displayed in REPSA experiments. For REPSA round 4, the test DNA was uncut in both control and experimental lanes. However, for REPSA round 7, the red-labeled test DNA was cleaved in the control lane, and a portion of it was uncleaved in the experimental third lane. These results indicate that the cleavage reactions in REPSA were selecting for sequences preferred and protected by TTHB099.

Furthermore, EMSA was performed using DNA from REPSA Rounds 1 and 7 to qualify the affinity of TTHB099 for each selection. Ten-fold dilutions of TTHB099

protein interacting with Round 1 DNAs did not show any visible protein-DNA complex formation, indicating that the protein does not bind to the majority of the sequences (Figure 6B, left). However, TTHB099 titrations with Round 7 DNAs displayed an increasing protein-DNA complex formation, represented by the increasing intensity in the mobility shift (Figure 6B, right). These results indicated that a substantial portion of the selected sequences in Round 7 contained stable TTHB099 binding sites.

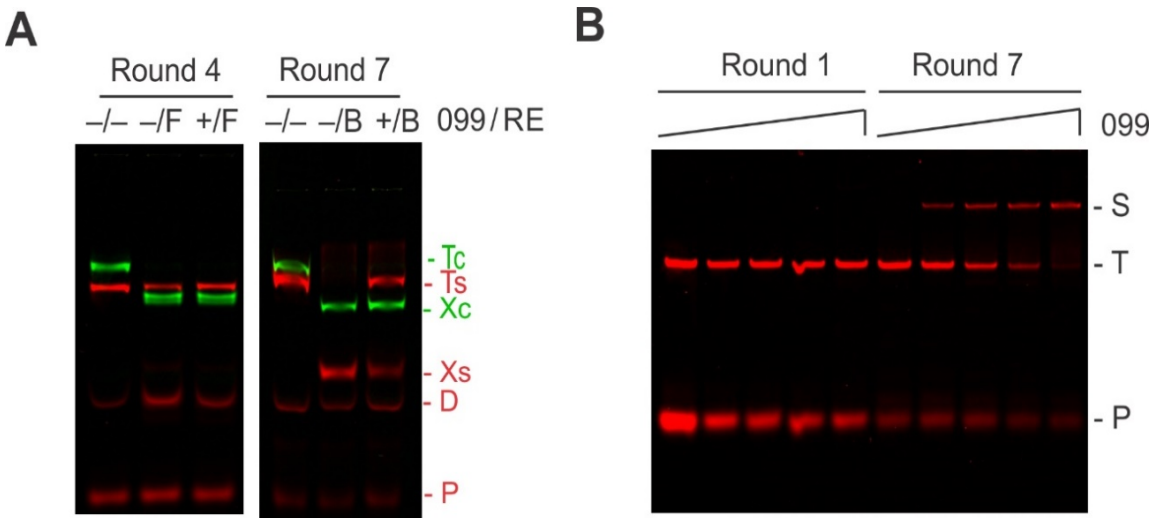

Figure 6. Validation of TTHB099-binding DNA sequences. (**A**) Shown are IR fluorescence images of restriction endonuclease protection assays made with DNA from Round 4 and 7 of REPSA selection. The presence (+) or absence (–) of TTHB099 and IISRE FokI (F) or BpmI (B) are indicated above each lane. The electrophoretic mobility of the intact (T) and cleaved (X) IRD8-labeled REPSAis control DNA (green), IRD7-labeled ST2R24 selection template (red), primer-dimers (D), as well as the IRD7_ST2R primer (P) are indicated at the right of the figure and color-coded to match the fluorescently labeled DNA present. (**B**) Shown are IR fluorescence images of electrophoretic mobility shift assays made with DNA mixtures obtained from Round 1 (left lanes) and Round 7 (right lanes) of REPSA selection incubated with increasing concentrations of TTHB099 protein (from left to right: 0, 5.06, 50.6, 506, and 5,060 nM TTHB099). The electrophoretic mobility of a single protein-DNA complex (S) as well as the uncomplexed ST2R24 selection template (T) and IRD7_ST2R primer (P) are indicated at the right of the figure.

Given the promising data obtained from the REPA and EMSA validations, massively parallel sequencing was performed on fusion libraries synthesized from Round 7 REPSA-selected DNAs. In this example, the ion semiconductor sequencing run yielded

6,921,164 total bases, 6,169,384 ≥ Q20, resulting in 120,585 reads of 57-bp mean length for the Round 7 DNA. Further analysis in Sequencing1.java refined individual sequences to 8,212 reads saved in fastq format. The MEME output displayed the best 23-mer nonpalindromic motif with an E-value of 2.4×10^{-2234} (Figure 7A) and the best 16-mer palindromic motif with an E-value of 2.4×10^{-2871} (Figure 7B). These statistically significant results indicate that the identified motifs are likely consensus sequences for the TTHB099 transcription factor. Noting that the nonpalindromic sequence logo is an extended version of the palindromic one, and because bacterial TFs tend to bind DNA as dimers, it was postulated that the palindromic logo is a better representation of the TTHB099 consensus DNA-binding sequence. Following that hypothesis, the 16-mer sequence 5'–TGTATTCTAGAATACA–3' was incorporated into an ST2 background, yielding the probe ST2_099.

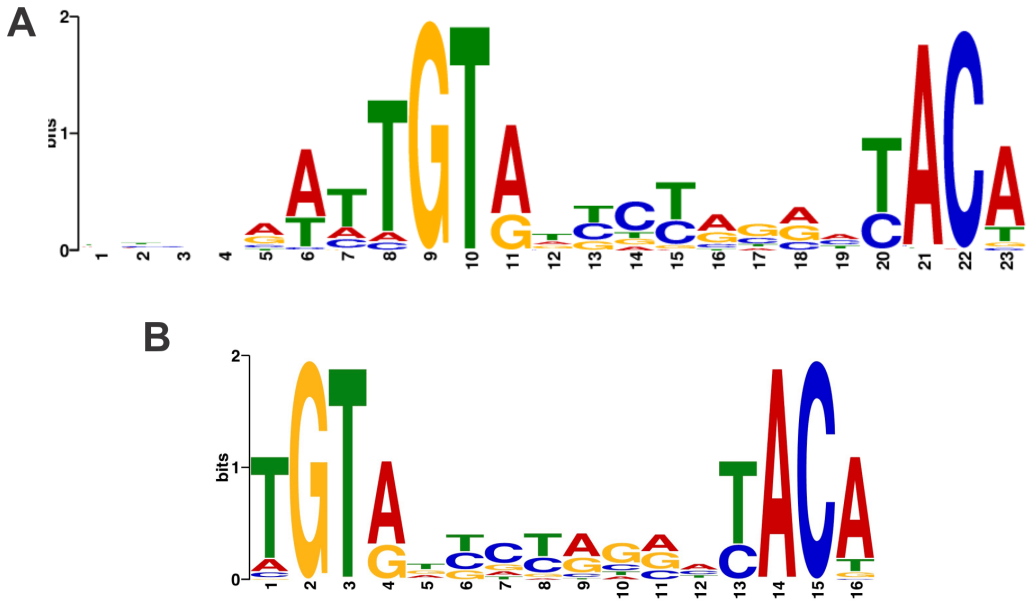

Figure 7. TTHB099-binding motifs. Sequence logos were determined using MEME software with an input of 1,000 Round 7 DNA sequences. (**A**) MEME performed with no filters. (**B**) MEME performed using a palindromic filter.

3.3 Characterization of the TTHB099 DNA-binding Motif

A fixed concentration of IRD7-labeled ST2_099 was incubated with increasing concentrations of purified TTHB099 protein to permit specific binding, and the resulting products analyzed by EMSA (Figure 8). We found that the TTHB099-ST2_099 complex exhibited similar electrophoretic mobility as observed with the TTHB099-Round 7 DNA complex (Figure 6B, left), suggesting that most Round 7 DNA contained the palindromic sequence. Indeed, this was found in our MEME results, where the palindromic sequence was present in 899/1,000 sites while the nonpalindromic was found in only 638/1,000 sites. Quantitative densitometry analysis of the fourth lane bands' intensities (Table 2) gave an approximate dissociation constant (K_D) of 4.5 nM.

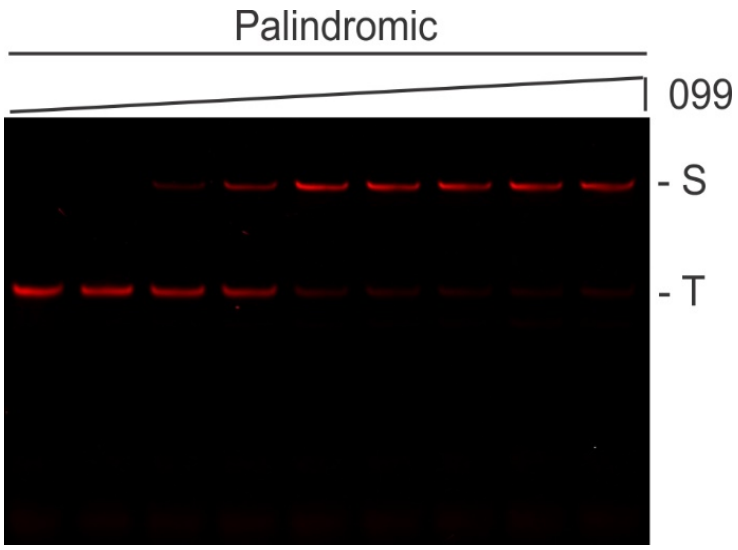

Figure 8. EMSA analysis of TTHB099 binding to its palindromic consensus sequence. Shown is an IR fluorescence image of IRD700-labeled ST2_099 incubated with 0, 0, 0.66, 1.32, 2.64, 5.27, 10.5, 21.1, or 42.2 nM TTHB099 protein. (S) Protein-DNA complex, and (T) uncomplexed DNA.

Table 2. EMSA quantification data.

Lane	[099] nM	Intensity S	Intensity T

1	0	BK	11,600,000
2	0	BK	10,700,000
3	0.66	2,040,000	8,270,000
4	1.32	4,680,000	7,080,000
5	2.64	9,540,000	2,280,000
6	5.27	8,540,000	1,850,000
7	10.5	8,090,000	1,450,000
8	21.1	8,530,000	1,160,000
9	42.2	7,820,000	1,520,000

(BK) Background noise due to the intensity being lower than the standard used by the LI-COR Odyssey Imager.

Following EMSA validation and K_D determination, a more sensitive technique such as BLI was used to characterize the binding affinity of TTHB099 to the palindromic ST2_099 sequence. This innovative approach measures *in vitro* real-time interactions between macromolecules, including proteins and nucleic acids. Our BLI analysis involved a biotinylated consensus sequence, ST2_099, affixed to streptavidin sensors interacting with increasing TTHB099 protein concentrations in solution. This assay provided a qualitative observation of protein-DNA association and dissociation kinetics (Figure 9A). The most substantial interactions were observed for the highest concentrations of TTHB099 (450 nM [red] and 150 nM [green]). An arbitrary DNA sequence, ST2_REPSAis, was tested as a control DNA (Figure 9B). It demonstrated binding interactions that were below our experimental detection levels, consistent with a low TTHB099-REPSAis affinity. Another outcome of this study was the quantitative evaluation of the TTHB099-consensus binding affinity. Least squares regression analysis of the association and dissociation rates were calculated with GraphPad Prism 8. From those rates, a dissociation constant was produced. TTHB099 interacting with its consensus sequence had a K_D of 2.214 nM with an R^2 value of 0.9883.

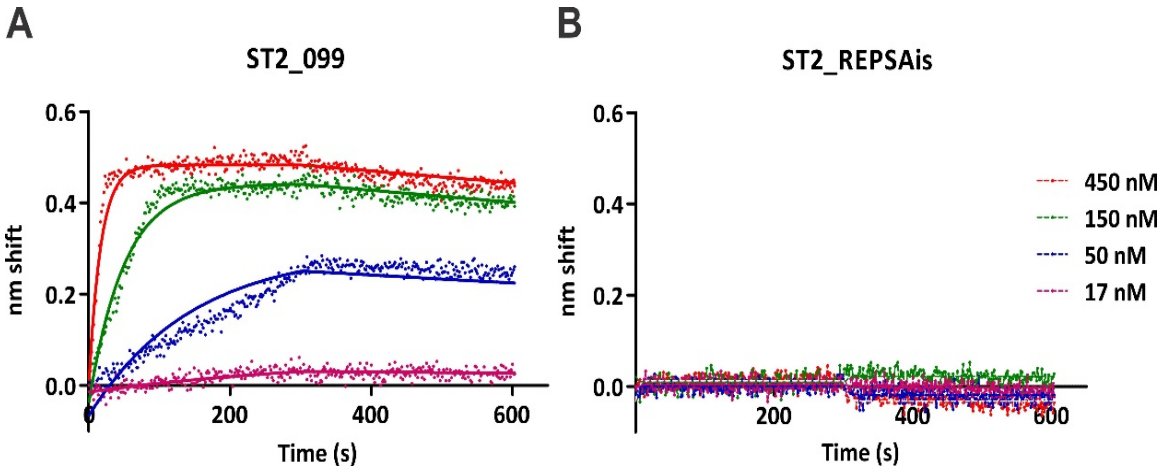

Figure 9. Biolayer interferometry analysis of TTHB099 binding to DNA. Shown are raw traces (dots) and best-fit lines of TTHB099 binding to (**A**) ST2_099 consensus DNA and (**B**) ST2_REPSAis control DNA TTHB099. Concentrations investigated include 450 nM (red), 150 nM (green), 50 nM (blue), and 17 nM (magenta).

Further characterization of TTHB099-DNA binding was made using selected point mutations of its consensus sequence and BLI. Binding kinetics data, including association rate (k_{on}), dissociation rate (k_{off}), and the dissociation constant, were derived for each of the mutated sequences and displayed in Table 3. As observed with the m2 mutant, a single change in a highly conserved nucleotide of the consensus sequence affects the binding affinity by 15-fold. Even point mutations of less conserved positions (e.g., m5) decreased the affinity by 2-fold. These data suggest that the TTHB099 binding to DNA is highly sequence-specific. Additionally, the nanomolar dissociation constant we observed indicates that our consensus sequence is a good representation of the native TTHB099s' preferred sequences in *T. thermophilus* HB8. Notably, TTHB099-DNA binding is not affected by the absence or presence of the second messenger 3', 5'cAMP, unlike its archetype protein CRP_{Ec} [67].

Table 3. TTHB099-DNA binding parameters for consensus and mutant sequences.

Name	Sequence	k_{on} ($M^{-1}s^{-1}$)	k_{off} (s^{-1})	K_D (M)	R^2
wt	TGTATTCTAGAATACA	131308	2.907×10^{-4}	2.214×10^{-9}	0.9883
m1	**g**GTATTCTAGAATACA	120059	7.558×10^{-4}	6.295×10^{-9}	0.9895
m2	T**t**TATTCTAGAATACA	112773	3.785×10^{-3}	3.356×10^{-8}	0.9778
m3	TG**a**ATTCTAGAATACA	88146	1.221×10^{-3}	1.385×10^{-8}	0.9824
m4	TGT**c**TTCTAGAATACA	142953	1.366×10^{-3}	9.557×10^{-9}	0.9817
m5	TGTA**c**TCTAGAATACA	110766	5.379×10^{-4}	4.856×10^{-9}	0.9879
m6	TGTAT**a**CTAGAATACA	125945	7.064×10^{-4}	5.608×10^{-9}	0.9794
m7	TGTATT**t**TAGAATACA	119827	6.978×10^{-4}	5.823×10^{-9}	0.9805
m8	TGTATTC**a**AGAATACA	115299	7.848×10^{-4}	6.807×10^{-9}	0.9840
wt + cAMP	TGTATTCTAGAATACA	214759	4.780×10^{-4}	2.226×10^{-9}	0.9231

(Sequence) Lowercase nucleotides indicate a mutation from the TTHB099 consensus sequence (wt). (wt + cAMP) Binding reactions performed with the consensus sequence in the presence of 100 nM 3',5'cAMP.

3.4 Genome-wide Mapping of the TTHB099 DNA-binding Motif

Following sequencing results, the MEME derived consensus sequence was entered into a FIMO analysis that revealed 78 motif occurrences with a p-value of less than 0.0001. The top 25 results with p-values $\leq 3.95 \times 10^{-5}$ are shown in Table 4. The locations of these 25 sequences relative to the TSS of their proximally downstream genes were determined using the KEGG database. Sixteen of these sites were situated within the –200 to +20 nucleotide region most common for transcription activator binding in bacteria. Furthermore, their proximally downstream genes were the first of their operons or single transcriptional units, making these sites stronger candidates for TF regulation. The other nine sites were omitted from further analysis because they were located further downstream, inside open reading frames, or, as in the case of TTHC003, too far upstream (–666 nucleotides).

Table 4. TTHB099-consensus sequences mapped in the genome of *T. thermophilus* HB8.

Start	End	*P*-value	*Q*-value	Sequence	Loc	Gene	Op
81408	81423	4.03×10^{-6}	1	AGTAAACTAAAACACA	+1	*TTHA0081*	1/3
				TGTGTTTTAGTTTACT	–48	*TTHA0080*	S
32704	32719	5.82×10^{-6}	1	TGTGTACGAAATTACA	+434	*TTHA0030*	1/2
472203	472218	7.74×10^{-6}	1	TGTATCTTGAAAAACA	–26	*TTHA0507*	S
				TGTTTTTCAAGATACA	–56	*TTHA0506*	S
130005	130020	1.01×10^{-5}	1	TTTATTCTCCCTTACA	–10	*TTHA0133*	1/2
				TGTAAGGGAGAATAAA	–3	*TTHA0132*	S
1506	1521	1.23×10^{-5}	1	AGTGAGATAACTCACA	–666	*TTHC003*	1/3
				TGTGAGTTATCTCACT	+627	*TTHC002*	S
79627	79642	1.30×10^{-5}	1	TGTGGTCCAGGCTACC	–78	*TTHB089*	1/3
				GGTAGCCTGGACCACA	–162	*TTHB088*	S
615132	615147	1.46×10^{-5}	1	GGTAGCCAGGGATACA	+909	*TTHA0647*	4/4
1715061	1715076	1.65×10^{-5}	1	TGTAGGCCAGGCCACG	–33	*TTHA1833*	1/2
609145	609160	1.83×10^{-5}	1	CGTGTCCCTGAACACA	+790	*TTHA0641*	2/4
614143	614158	2.12×10^{-5}	1	TGTGCCTTTGGCCACA	+326	*TTHA0645*	1/3

1794923	1794938	2.33 x 10^{-5}	1	GGTATGCTCAAGTACA	+13	*TTHA1912*	1/2
				TGTACTTGAGCATACC	–19	*TTHA1911*	1/4
1272	1287	2.61 x 10^{-5}	1	TGTAGCCCAGGCCAAA	+239	*TTHB003*	S
				TTTGGCCTGGGCTACA	+536	*TTHB004*	4/4
199120	199135	2.90 x 10^{-5}	1	TGTGGCGTATAACAAA	–17	*TTHA0202*	S
				TTTGTTATACGCCACA	–103	*TTHA0201*	S
357035	357050	3.43 x 10^{-5}	1	AGTGATGTAAACTAAA	–26	*TTHA0374*	S
314103	314118	3.67 x 10^{-5}	1	TGTGTTGCAGGACCCA	+58	*TTHA0326*	2/11
1540358	1540373	3.95 x 10^{-5}	1	TGTAGCTTCCCATACC	–67	*TTHA1627*	S
				GGTATGGGAAGCTACA	+13	*TTHA1626*	S

(*P*-value) The probability of a random equally long sequence matching that position of the sequence with an as good or better score. (*Q*-value) False discovery rate if the occurrence is accepted as significant. (Loc) Location of the TTHB099-binding site relative to the start site of translation. (Gene) Proximal gene downstream of the TTHB099 consensus sequence. (Op) Gene position within the postulated operon. (S) No operon, single transcriptional unit.

To better ascertain a potential role for TTHB099 to regulate transcription, all 16 sequences selected above were analyzed for potential core promoter elements. Sequences ± 200 bp upstream and downstream of the FIMO identified TTHB099-binding sites were evaluated in SoftBerry BPROM (Figure 6). Many sequences (9/16) contained a TTHB099-consensus site overlapping with at least one promoter element (–35 box, –10 box, +1 start site). Those included *TTHA0081/80, TTHA0507, TTHA0133, TTHA1833, TTHA1912, TTHA0202, TTHA0374,* and *TTHA1627*. Three of the TTHB099-binding sequences*, TTHA0506, TTHB089*, and *TTHA0201,* were located upstream of the nearby –35 box. Conversely, *TTHB088* and *TTHA1626* had their putative TTHB099-binding sequences located downstream of the postulated promoter elements. There were no identified promoter elements near *TTHA0132* and *TTHA1911.* It is not clear why BPROM was unable to identify any core promoter elements, but limitations could arise from a potential difference between core promoter elements in *E. coli*, the model organism used by BPROM, and those of *T. thermophilus* HB8.

>TTHA0080, complement(81208 ... 81623)
GCGCTCCGCGTCGGAGAGGACCTCGTAGGCGTCGTGCCAGCCCTCCGGGCCCACCACCTT
GTCCAGGCGGTCCAACACGGTGCGGGCGTCCACGTAGGGGACCACCAAGGCCCGCTTCTT
GTCCCGGGAGAGGGCTTCCACGCGCCACTGCACCTCCCCCGGGGGAAAGGGTTCGGCCAG
TTTCCGCCAGACTTCGTCCATGTGGTTTTAGTTTACTTAGGTTGCTCTCACCCCAAAGCC
TTGGGGGAAGGCGAAGATGGGGGCATGAAGCGGTGGCTGGCGTTCCTTCCCTTCCTGGCC
CTGGCCTGGGCTTTGGAGCTCAGGGTCACCGCCTCCTTGGTGGTGGACCTCTTCCCCCAG
GCGGTGGTGGTGGAGCGGGTTACCGAGCCCCAGGGGATCGTGGTGGTTTACCAGGC

>TTHA0506, complement(472003 .. 472418)
GCCCGCCTGGGCCAGGGCCCTGAGGAGGCGGTAGAGGGTGCTCTTGGCGAGGCCCACCCG
TTCGGCCAAGGGGCCCAAGGGGCTTTCCCCCGCCTGGGCCAGGGCCTCGAGGACCCGAAG
CCCCCTCTCCAGGGTCTTCACCGCCTGGGGCGGCTTCTCCCGAGGACGCGCCATGCCGCT
TAGGGTAACGGGGGCGGCCCTGTTTTTCAAGATACAAAAATCTTTTTGCTTCTTGACAA
TCCCGCCCCGCCTCCCGTAAGCTCGGACCACCATGAAGGGCGTGGAGATCCGGAAAGACC
ACCCCCTCCTGAAGGAGGTCCTGACGGAGGAGGCCCTGAGGTTCGTGGTGGCGCTGCACC
GGGAGTTCAACCCGGTGCGCAAGGCCCTCCTGGAGCGGCGTCAAGCGCTTTGGGAG

>TTHA0132, complement(129522 .. 130201)
TGGGGATGGCCGTGGCCCCAAGGGCCTCCACCTCCGAGGCCACCCCCGTGGCGAGCTCCA
CGTCGGGGTCCACGGCGATGACCGTGGCCCCGTTGCGCCCGTACCCGTGGGCGATGGCCC
GCCCGAACCCCCGGCCCGCGCCCGTGACCATGACGATCTTCCCCTCGAGGCCCAGAAGGT
CCCGTGACATCACGGCCCATTGTAAGGGAGAATAAAGCCATGGCGCGCATCCGGGTGGTC
CAAGGGGACATCACCGAGTTCCAAGGGGACGCCATCGTCAACGCCGCCAACAACTACCTG
AAGCTCGGGGCCGGGGTGGCGGGGGCGATCCTGAGGAAGGGCGGCCCCTCCATCCAGGAG
GAGTGCGACCGCATCGGCAAGATCCGGGTGGGGGAGGCGGCGGTCACGGGGGCGGG

>TTHB088, complement(79427 ... 79842)
TTCACCAGGACGTCCACCGCCGGGGCGTCGGGGGAGAGGTGGGCCACCCGCACCATGGCG
CCTTGGCCCAGGGCCAGGCCGGCCAGGGCCGCCAGAACCAGAACAAAAAGGCCTCGTTTC
ATCTTTTCACCTCCACGGGAAAAGCCTAGAGGGAGGCCTGCCCGTCAAAATGGGCGCAGG
CCACATAAACCTCCCGCCAAGGTAGCCTGGACCACACCCAGGGTGAGGGGGAGCACATTC
TCGGGGGACCTTCGGCCCTAGCATCCTCCCAAAGGAGGTAAGGGCATGGACCGCAGGCGT
TTTCTCACCGGTGCGGGGCTTTTTTTGGCGGCGGGAGGCCTTCCCTTGGGCCGGCCCAG
GGGCGCGCGCCCAAGGGGGTGAACGGGGGCGGCTTTTACCGCTTCCGGGTAGGGGG

>TTHA1833, (1714861 .. 1715276)
AACCATCGTTCCCCTGAGGCAGGCCCTGGGCTTTAGGATCCTCGGGGCCTACTGGCTTTC
CGAGCGGGAGTTCCTCTGGTTCGTGGCCCACGAGGACTTTGAGGAGGCGGAGAGAGCTTA
CTACGCCCACCCCGAAAGGCAGAAGGTGGACCCCAGGGCGTACCTGGAGGCGGTGGAAAC
CCGTTTCGTGGAACGCCTTCTGTAGGCCAGGCCACGCCCCTGGCCCCGCCTTGGGTAGC
CTCGGAGGGATGGAGCTTTTCCTCCTCGTCCTCCGCAACCTCCTGGCCCGGCCCGTCCGG
AGCCTCCTCACCCTGCTCGGGGTCCTGGTGGCCACGGCGAGCATGGTCCTCTTCCTCTCC
TTCGGGGAGGGCCTTAGGCGGGCCCTCTTCCAGGAGCTCTCCCGGGTGGGCCCCGC

>TTHA1912, (1794726 .. 1795150)
CGATGGAGACCCTTTCCGGGTAGCGGGGGGTGGCCTCCAGGTACTCCAGGCGCTTGAAGA
AGCTCCCGGCGATGGAGTCCACCACCATCACCTGGTCCACCTCCACCACCACGAGCTCCC
CCGCCCGCACGGGCCTTCCCACCTTGTGGGAGAGGATCTTTTCCGCTAGCGTCTGTCCCA
CGGACACCTCCTATACTGAGGGTATGCTCAAGTACCACCGCCCTCCTCTACCCGGACCCGG
AGACCCCAGGGGTCTGGATCGCCGAGTTTCCCGCGGTGCCCCAAGCCCACTCCTTCGGCC
AAAGCCCCGAGGAAGCCTTGGCGCGGGCCAAAGAAGCCCTGGAGCTCGTCCTGGCCTATC
TGAAAACCGAGGGGCGCCCCCTTCCCCAGGACGTACAAGCGGTAGAGGTAGGTGTG

>TTHA0202, (198920 .. 199335)
GGGCGTAGCTGGGAAGCCCCGGGCTAACGTCCACCTCCACGGTGACGGGAACCGCGTCCA
GGCCGAAGAGGGCGTAGCTTCGCACCTGGGCCAGCATGGAGAGAGTTTATCACAGCGCTG
TTAGTTCCACCCAAGGTGGGCGTTTCGTGAGCAGAGGCGAAAACTGCCTTATCATGGGGC
CAGATGGCGCGCGAGCTCCATGTGGCGTATAACAAAATGCGCCGCGCCCTGGAGGAGCGC
TTGGGCCTCCTCCGCCGCCTCGGGGGAATGGACCTCCGCTTGATCCAAGTGGCCAACGAG
GAGTGGCTCTACATGCTCCAGGAGGACACCCGCAACTCCCTGGCCATAGAGGGCTACTTC
ACCACGGAGCGGGAGCTACGGGAGGTGCTTAGGGGACGCAAGGGGGCGGCGGAGGT

>TTHA1626, complement(1540158 .. 1540573)
CCGTGCCCGACTGGGGCAGGCCCTCCACCCCCAAGGTGGCGTCCAAGGGGGTGGGGGCGA
GGGCGAGGAGGAAGGGGAAGAGGTCCTCCAAGGCTCCTCCCAGGCCCAGGCCGTCCACGA
AGAGGAAGAGCACCTTTCCATCCTAAGGGGGGACATTTGCCACAGGGGGATAGAGGTACC
CTGAGCTTAGGAGGTGATGGGGTATGGGAAGCTACAACCCGCTGGTCTTCGTTCTAGGCC
TGGTCACGGCGGCCGGGGTCTCGGGGGTGGCCTACTTGCTCGCCGTGGCCCGGGGTGGGG
ACGAGAAGGCCCTGGGGCGGCTTTATGGCCCCCTCTTCTTCACCCTGGGGGTCTTCTCCC
TGGGGGCGGTGGCCCAGCTCTACTGGACCAACTGGGCGGGCCGTCCGGTGCCCCAG

>TTHA0081, (81208 .. 81623)
GCCTGGTAAACCACCACGATCCCCTGGGGCTCGGTAACCCGCTCCACCACCACCGCCTGG
GGGAAGAGGTCCACCACCAAGGAGGCGGTGACCCTGAGCTCCAAAGCCCAGGCCAGGGCC
AGGAAGGGAAGGAACGCCAGCCACCGCTTCATGCCCCCATCTTCGCCTTCCCCCAAGGCT
TTGGGGTGAGAGCAACCTAAAGTAAACTAAAACACATGGACGAAGTCTGGCGGAAACTGG
CCGAACCCTTTCCCCCGGGGGAGGTGCAGTGGCGCGTGGAAGCCCTCTCCCGGGACAAGA
AGCGGGCCTTGGTGGTCCCCTACGTGGAGCGCCGCACCGTGTTGGACCGCCTGGACAAGG
TGGTGGGCCCGGAGGGCTGGCACGACGCCTACGAGGTCCTCTCCGACGCGGAGCGC

>TTHA0507, (472003 ... 472418)
CTCCCAAAGCGCTTGACGCCGCTCCAGGAGGGCCTTGCGCACCGGGTTGAACTCCCGGTG
CAGCGCCACCACGAACCTCAGGGCCTCCTCCGTCAGGACCTCCTTCAGGAGGGGGTGGTC
TTTCCGGATCTCCACGCCCTTCATGGTGGTCCGAGCTTACGGGAGGCGGGGCGGGATTGT
CAAGAAGCAAAAAGATTTTTTGTATCTTGAAAAACAGGGCCGCCCCCGTTACCCTAAGCG
GCATGGCGCGTCCTCGGGAGAAGCCGCCCCAGGCGGTGAAGACCCTGGAGAGGGGGCTTC
GGGTCCTCGAGGCCCTGGCCCAGGCGGGGGAAAGCCCCTTGGGCCCCTTGGCCGAACGGG
TGGGCCTCGCCAAGAGCACCCTCTACCGCCTCCTCAGGGCCCTGGCCCAGGCGGGC

>TTHA0133, (129831 .. 130938)
CCCGCCCCCGTGACCGCCGCCTCCCCCACCCGGATCTTGCCGATGCGGTCGCACTCCTCC
TGGATGGAGGGGCCGCCCTTCCTCAGGATCGCCCCCGCCACCCCGGCCCCGAGCTTCAGG
TAGTTGTTGGCGGCGTTGACGATGGCGTCCCCTTGGAACTCGGTGATGTCCCCTTGGACC
ACCCGGATGCGCGCCATGGCTTTATTCTCCCTTACAATGGGCCGTGATGTCACGGGACCT
TCTGGGCCTCGAGGGGAAGATCGTCATGGTCACGGGCGCGGGCCGGGGGTTCGGGCGGGC
CATCGCCCACGGGTACGGGCGCAACGGGGCCACGGTCATCGCCGTGGACCCCGACGTGGA
GCTCGCCACGGGGGTGGCCTCGGAGGTGGAGGCCCTTGGGGCCACGGCCATCCCCA

>TTHB089, (79427 ... 79842)
CCCCCTACCCGGAAGCGGTAAAAGCCGCCCCCGTTCACCCCCTTGGGCGCGCGCCCCTGG
GCCCGGCCCAAGGGAAGGCCTCCCGCCGCCAAAAAAAGCCCCGCACCGGTGAGAAAACGC
CTGCGGTCCATGCCCTTACCTCCTTTGGGAGGATGCTAGGGCCGAAGGTCCCCCGAGAAT
GTGCTCCCCCTCACCCTGGGTGTGGTCCAGGCTACCTTGGCGGGAGGTTTATGTGGCCTG
CGCCCATTTTGACGGGCAGGCCTCCCTCTAGGCTTTTCCCGTGGAGGTGAAAAGATGAAA
CGAGGCCTTTTTGTTCTGGTTCTGGCGGCCCTGGCCGGCCTGGCCCTGGGCCAAGGCGCC
ATGGTGCGGGTGGCCCACCTCTCCCCCGACGCCCCGGCGGTGGACGTCCTGGTGAA

>TTHA1911, complement(1794726 .. 1795150)
CACACCTACCTCTACCGCTTGTACGTCCTGGGGAAGGGGGCGCCCCTCGGTTTTCAGATA
GGCCAGGACGAGCTCCAGGGCTTCTTTGGCCCGCGCCAAGGCTTCCTCGGGGCTTTGGCC
GAAGGAGTGGGCTTGGGGCACCGCGGGAAACTCGGCGATCCAGACCCCTGGGGTCTCCGG
GTCCGGGTAGAGGAGGGCGGTGTACTTGAGCATACCCTCAGTATAGGAGGTGTCCGTGGG
ACAGACGCTAGCGGAAAAGATCCTCTCCCACAAGGTGGGAAGGCCCGTGCGGGCGGGGGA
GCTCGTGGTGGTGGAGGTGGACCAGGTGATGGTGGTGGACTCCATCGCCGGGAGCTTCTT
CAAGCGCCTGGAGTACCTGGAGGCCACCCCCCGCTACCCGGAAAGGGTCTCCATCG

>TTHA0201, complement(198920 .. 199335)
ACCTCCGCCGCCCCCTTGCGTCCCCTAAGCACCTCCCGTAGCTCCCGCTCCGTGGTGAAG
TAGCCCTCTATGGCCAGGGAGTTGCGGGTGTCCTCCTGGAGCATGTAGAGCCACTCCTCG
TTGGCCACTTGGATCAAGCGGAGGTCCATTCCCCCGAGGCGGCGGAGGAGGCCCAAGCGC
TCCTCCAGGGCGCGGCGCATTTTGTTATACGCCACATGGAGCTCGCGCGCCATCTGGCCC
CATGATAAGGCAGTTTTCGCCTCTGCTCACGAAACGCCCACCTTGGGTGGAACTAACAGC
GCTGTGATAAACTCTCTCCATGCTGGCCCAGGTGCGAAGCTACGCCCTCTTCGGCCTGGA
CGCGGTTCCCGTCACCGTGGAGGTGGACGTTAGCCCGGGGCTTCCCAGCTACGCCC

>TTHA0374, (356835 .. 357250)
CAACAACGTGGATCCCGAGCGGGACGCCCGGGTGATGCCGGGGGTGGAGGGGCCGGTTTT
GGTCCTGGACGGCACGAGGAAGCTCCCCGAGGAGGGCTTCCCCAGGGTCTGGCCCGAGAG
GATCCGGATGGACCCCAAGGTGAAGGCCTTGGTGGAGGCCCGGTGGGCGGAGTACGGCCT
GGGCTGGACAACGGTGGGTGAGTGATGTAAACTAAAAGAGGTTTAGTGCGAAGGTGATAT
TTATGGGCTTACACGTCCTCGGCGTGAACGCATCGGCTAGGACGGACGGGTTTACGGCGG
AGCTTTTGGACGAGGTTTTGGAGGCGGCCAGGCGCAAGGGGGCGACCACCGAGCGCCTGG
ATTTGGTGCGGCACCCCTTTCCCCTCTGCGCCGGCAACTACTCCGTGGACCCCGCT

>TTHA1627, (1540158 .. 1540573)
CTGGGGCACCGGACGGCCCGCCCAGTTGGTCCAGTAGAGCTGGGCCACCGCCCCCAGGGA
GAAGACCCCCAGGGTGAAGAAGAGGGGGCCATAAAGCCGCCCCAGGGCCTTCTCGTCCCC
ACCCCGGGCCACGGCGAGCAAGTAGGCCACCCCCGAGACCCCGGCCGCCGTGACCAGGCC
TAGAACGAAGACCAGCGGGTTGTACTTCCCATACCCCATCACCTCCTAAGCTCAGGGTA
CCTCTATCCCCCTGTGGCAAATGTCCCCCCTTAGGATGGAAAGGTGCTCTTCCTCTTCGT
GGACGGCCTGGGCCTGGGAGGAGCCTTGGAGGACCTCTTCCCCTTCCTCCTCGCCCTCGC
CCCCACCCCCTTGGACGCCACCTTGGGGGTGGAGGGCCTGCCCCAGTCGGGCACGG

Figure 10. Promoter predictions of sequences potentially regulated by TTHB099 within the *T. thermophilus* HB8 genome. Shown are ±200 bp sequences from the TSS of the genes identified through FIMO (see Table 4). Blue nucleotides represent the longest open reading frames with a downstream orientation relative to the TTHB099 binding site; Green nucleotides indicate open reading frames with the opposite orientation; Black nucleotides imply intergenic regions. Potential promoter elements (–35 and –10 boxes, +1 start site of transcription) are indicated with cyan highlighting; TTHB099-binding sites are indicated with yellow highlighting; Overlapping TTHB099-binding and core promoter elements are indicated by green highlighting.

3.5 Validation of Potential TTHB099-regulated Genes

Apart from analyzing the locations of the binding sequences concerning the transcription start site, as well as their positions regarding promoters, we investigated the affinity of TTHB099 protein for the selected sequences. To better understand how TTHB099 regulates genes identified through FIMO, all 16 sequences underwent binding kinetics analysis via BLI. As some TTHB099 binding sites are shared by two bidirectional promoters, only nine unique sequences were synthesized into biotinylated double-stranded oligonucleotides. Binding reactions containing four different concentrations of TTHB099 (450, 150, 50, and 17 nM) were tested against each binding site probe (Table 5).

The strongest binding was observed for *TTHA1833* and *TTHB088/89,* with K_D values below 10 nM. The genes with binding affinities between 10–100 nM were *TTHA1911/12*, *TTHA0506/07*, and *TTHA0080/81* in increasing order. *TTHA1626/27*, *TTHA0132/33*, and *TTHA0201/02* displayed the weakest binding, with K_{Ds} >100 nM, while binding to *TTHA0374* could not be detected under our experimental conditions. These binding parameters do not follow the sequence order defined by FIMO, suggesting that there could be other factors in effect that are not considered in this *in vitro* analysis.

Table 5. Binding kinetics parameters of TTHB099 to potential gene promoter elements.

Gene	Sequence	k_{on} ($M^{-1}s^{-1}$)	k_{off} (s^{-1})	K_D (M)	R^2
TTHA0080/81	TGTGTTTTAGTTTACT	122852	1.145×10^{-2}	9.322×10^{-8}	0.9817
TTHA0506/07	TGTTTTTCAAGATACA	164971	1.280×10^{-2}	7.762×10^{-8}	0.9718
TTHA0132/33	TGTAAGGGAGAATAAA	96736	2.140×10^{-2}	2.212×10^{-7}	0.9687
TTHB088/89	GGTAGCCTGGACCACA	214153	7.163×10^{-4}	3.345×10^{-9}	0.9805
TTHA1833	TGTAGGCCAGGCCACG	332611	1.013×10^{-3}	3.046×10^{-9}	0.9757
TTHA1911/12	TGTACTTGAGCATACC	136294	8.938×10^{-3}	6.558×10^{-8}	0.9806

TTHA0201/02	TTTGTTATACGCCACA	57231	0.04464	7.801×10^{-7}	0.9596
TTHA0374	AGTGATGTAAACTAAA	–	–	–	–
TTHA1626/27	GGTATGGGAAGCTACA	126605	1.291×10^{-2}	1.020×10^{-7}	0.9759

(TTHA0080/81) Two bidirectional promoters share a common TTHB099-binding site. (–) No apparent binding.

Further validation of TTHB099 involvement in the transcriptional regulation of these genes as well as their operons was sought through the analysis of prior DNA microarray studies, publicly available through the National Center for Biotechnology Information Gene Expression Omnibus (NCBI GEO) [63]. A GEO2R comparison of expression profile data from sets of TTHB099-deficient and wild type strains (SuperSeries GSE21875) was used to determine if the absence of *TTHB099* produced any substantial changes in the expression of the FIMO-identified genes and their operons. Of these genes, only *TTHA1626* displayed a substantially increased expression with a logFC of 2.62. The remainder of the 15 genes had only small, non-significant changes, as shown in Table 6. Likewise, individual genes within their respective operons did not seem to have any significant changes.

Table 6. Expression profile data of the FIMO identified operons in a TTHB099-deficient strain of *T. thermophilus* HB8.

Operon	Gene	Role	LogFC	Adj *P*-value
S	*TTHA0080*	hypothetical protein	0.851	0.0268
1	*TTHA0081*	hypothetical protein	–0.202	0.421
2	*TTHA0082*	phosphoesterase	–0.176	0.463
3	*TTHA0083*	dimethyladenosine transferase	–0.219	0.336
S	*TTHA0506*	malate synthase	–0.454	0.0983
S	*TTHA0507*	IclR family transcriptional regulator, acetate operon repressor	0.276	0.619
S	*TTHA0132*	hypothetical protein	0.872	0.0295

1	*TTHA0133*	Short-chain dehydrogenase/reductase family oxidoreductase	–0.211	0.674
2	*TTHA0134*	NrdR family transcriptional regulator	–0.328	0.350
S	*TTHB088*	Zn-dependent hydrolase	–0.386	0.653
1	*TTHB089*	hypothetical protein	–0.779	0.0451
2	*TTHB090*	hypothetical protein	–0.0653	0.955
3	*TTHB091*	hypothetical protein	–0.217	0.674
1	*TTHA1833*	ABC transporter permease	–0.294	0.287
2	*TTHA1834*	ABC transporter ATP-binding protein	–0.195	0.567
1	*TTHA1911*	3-isopropylmalate dehydratase large subunit	–0.817	0.0246
2	*TTHA1910*	homoaconitate hydratase small subunit	–1.14	0.0265
3	*TTHA1909*	hypothetical protein	–0.0793	0.790
4	*TTHA1908*	hypothetical protein	–0.0327	0.905
1	*TTHA1912*	hypothetical protein	0.353	0.154
2	*TTHA1913*	hypothetical protein	0.723	0.0284
S	*TTHA0201*	Mg^{2+} chelatase family protein	0.141	0.698
S	*TTHA0202*	hypothetical protein	0.454	0.0644
S	*TTHA0374*	hypothetical protein	0.687	0.0421
S	*TTHA1626*	hypothetical protein	2.62	2.10×10^{-3}
S	*TTHA1627*	hypothetical protein	–1.20	0.0960

(Operon) Numbers indicate positions of the genes within the operon. (S) Single transcriptional unit. (Role) The biological functions were identified using the KEGG database [47]. (LogFC) Log2-fold change between data obtained from TTHB099-deficient (accessions GSM530118/20/22) and wild-type (accessions GSM532194/5/6) *T. thermophilus* HB8 strains, SuperSeries GSE21875. (Adj. p-value) The p-value was obtained following multiple testing corrections using the default Benjamini and Hochberg false discovery rate method [68].

As an additional approach to better understand potential gene regulation by TTHB099, we investigated the postulated biological functions of these genes. Many were reported only as encoding hypothetical proteins, which is fairly common in *T. thermophilus*. Several encoded proteins may be involved in sugar metabolism (malate synthase, 3-isopropylmalate dehydratase), energy metabolism (3-isopropylmalate dehydratase large subunit, homoaconitate hydratase small subunit), transport, or others

(different pathways). Most interesting, two genes (*TTHA0134, TTHA0507*) are believed to encode transcriptional regulators. If so, their expression could complicate the identification of TTHB099 directly regulated genes by GEO2R.

Another analysis of the GEO2R data was focused on investigating the genes that were affected in the *TTHB099*-deficient strain (Table 7). These genes could be grouped into operons, suggesting that their expression was not affected by multiple-unrelated TFs, but rather a fundamental regulatory mechanism involving TTHB099. The upregulated genes, 75% (50/67), were involved in the electron transport chain (ETC) of oxidative phosphorylation as part of energy metabolism, carbohydrate metabolism, signaling and secretion, cofactor and vitamin metabolism, as well as others (Figure 12). The downregulated operons, 25% (17/67 genes), were related to ribosomal proteins, ion ABC transporters, and others (Figure 13). MEME analysis of the –300/+100 bp sequences upstream of each operon did not find our TTHB099 consensus sequence or reveal any additional binding motifs. Taken together, this suggests a complicated mechanism for the regulation of these genes that does not involve TTHB099 directly regulating their transcription.

Table 7. GEO2R analysis of the most affected genes in the *TTHB099*-deficient strain.

Operon	Gene	Role	LogFC	Adj. P-value
1	*TTHA1498*	Elongation Factor G	+ 4.384	2.07×10^{-4}
2	*TTHA1499*	MoxR-like protein	+ 5.067	7.03×10^{-5}
3	*TTHA1500*	Phosphoenolpyruvate Synthase	+ 5.231	7.03×10^{-5}
4	*TTHA1501*	Hemolysin III	+ 3.133	1.27×10^{-3}
5	*TTHA1502*	Response Regulator_two-component system, OmpR family	+ 1.087	9.51×10^{-3}
6	*TTHA1503*	Sensor Histidine Kinase	+ 0.369	2.46×10^{-1}
S	*TTHA1836*	Isocitrate lyase	+ 4.423	1.52×10^{-4}
1	*TTHA1838*	SufC protein, ATP-binding protein	–2.465	1.06×10^{-3}

2	*TTHA1839*	SufB protein, membrane protein	–2.593	9.53×10^{-4}
3	*TTHA1840*	SufD protein, membrane protein	–2.630	6.25×10^{-4}
4	*TTHA1841*	Dioxygenase ferredoxin subunit	–2.419	2.59×10^{-3}
1	*TTHA1133*	ba3-type cytochrome C oxidase polypeptide IIA	+ 1.311	4.37×10^{-2}
2	*TTHA1134*	ba3-type cytochrome C oxidase polypeptide II	+ 2.944	7.89×10^{-3}
3	*TTHA1135*	ba3-type cytochrome C oxidase polypeptide I	+ 4.269	1.27×10^{-3}
1	*TTHA1136*	hypothetical protein	+ 1.910	1.29×10^{-3}
2	*TTHA1137*	Major facilitator superfamily transporter	+ 2.300	9.53×10^{-4}
S	*TTHA0251*	Elongation factor Tu	–1.254	1.17×10^{-2}
1	*TTHA0250*	50S ribosomal protein L33	–1.139	8.04×10^{-3}
2	*TTHA0249*	Preprotein translocase subunit SecE	–0.997	9.18×10^{-3}
3	*TTHA0248*	Transcription antitermination protein NusG	–1.136	7.86×10^{-3}
1	*TTHA0247*	50S ribosomal protein L11	–2.378	1.27×10^{-3}
2	*TTHA0246*	50S ribosomal protein L1	–1.776	2.16×10^{-3}
1	*TTHA0084*	NADH-quinone oxidoreductase subunit 7	+ 1.083	8.73×10^{-3}
2	*TTHA0085*	NADH dehydrogenase subunit B	+ 1.005	2.41×10^{-2}
3	*TTHA0086*	NADH-quinone oxidoreductase subunit 5	+ 1.251	1.06×10^{-2}
4	*TTHA0087*	NADH-quinone oxidoreductase subunit 4	+ 1.255	6.43×10^{-3}
5	*TTHA0088*	NADH-quinone oxidoreductase subunit 2	+ 0.693	4.43×10^{-2}
6	*TTHA0089*	NADH-quinone oxidoreductase subunit 1	+ 1.249	4.68×10^{-3}
7	*TTHA0090*	NADH-quinone oxidoreductase subunit 3	+ 1.248	5.76×10^{-3}
8	*TTHA0091*	NADH-quinone oxidoreductase subunit 8	+ 1.490	3.62×10^{-3}
9	*TTHA0092*	NADH-quinone oxidoreductase subunit 9	+ 1.502	2.21×10^{-3}
10	*TTHA0093*	NADH-quinone oxidoreductase subunit 10	+ 1.626	6.84×10^{-3}
11	*TTHA0094*	NADH-quinone oxidoreductase subunit 11	+ 1.043	6.39×10^{-3}
12	*TTHA0095*	NADH-quinone oxidoreductase subunit 12	+ 1.492	2.85×10^{-3}
13	*TTHA0096*	NADH-quinone oxidoreductase subunit 13	+ 1.679	3.34×10^{-3}
14	*TTHA0097*	NADH-quinone oxidoreductase subunit 14	+ 1.509	2.84×10^{-3}
15	*TTHA0098*	arginyl-tRNA synthetase	+ 0.397	8.43×10^{-2}
16	*TTHA0099*	serine protease	+ 0.106	6.09×10^{-1}
17	*TTHA0100*	UDP-N-acetylmuramoylalanyl-D-glutamate--2,6-diaminopimelate ligase	+ 0.520	5.11×10^{-2}
S	*TTHA1626*	hypothetical protein	+ 2.616	2.10×10^{-3}
S	*TTHA1625*	Osmotically inducible protein OsmC	+ 1.206	3.65×10^{-3}
1	*TTHA1628*	Iron ABC transporter substrate-binding protein	–2.947	1.83×10^{-3}
2	*TTHA1629*	Iron ABC transporter permease	–2.344	1.68×10^{-3}
3	*TTHA1630*	Iron ABC transporter ATP-binding protein	–0.796	1.69×10^{-2}
4	*TTHA1631*	tRNA pseudouridine synthase A	–0.461	8.43×10^{-2}

S	*TTHA0135*	MutT/nudix family protein	–1.369	6.82×10^{-3}
1	*TTHA0206*	nicotinamide nucleotide transhydrogenase subunit alpha 1	+ 1.516	5.30×10^{-3}
2	*TTHA0207*	nicotinamide nucleotide transhydrogenase subunit alpha 2	+ 1.596	2.85×10^{-3}
3	*TTHA0208*	nicotinamide nucleotide transhydrogenase subunit beta	+ 1.647	2.10×10^{-3}
1	*TTHA0209*	50S ribosomal protein L10	–1.673	5.33×10^{-3}
2	*TTHA0210*	50S ribosomal protein L7/L12	–1.326	8.74×10^{-3}
1	*TTHB117*	putative type IV pilin	+ 1.125	4.09×10^{-2}
2	*TTHB118*	secretion system protein	+ 1.450	3.74×10^{-3}
3	*TTHB119*	prepilin-like protein	+ 1.429	5.85×10^{-3}
4	*TTHB120*	hypothetical protein	+ 2.250	1.27×10^{-3}
1	*TTHA1652*	maltose ABC transporter substrate-binding protein	+ 1.787	1.72×10^{-3}
2	*TTHA1651*	maltose ABC transporter permease	+ 2.154	1.17×10^{-3}
3	*TTHA1650*	maltose ABC transporter permease	+ 2.108	1.29×10^{-3}
1	*TTHB186*	putative transcriptional regulator	+ 3.377	2.59×10^{-3}
2	*TTHB187*	hypothetical protein	+ 2.036	7.58×10^{-3}
1	*TTHB188*	hypothetical protein	+ 1.215	9.19×10^{-3}
2	*TTHB189*	CRISPR-associated Cse2 family protein	+ 1.514	4.80×10^{-3}
3	*TTHB190*	hypothetical protein	+ 1.671	6.62×10^{-3}
4	*TTHB191*	hypothetical protein	+ 1.480	4.34×10^{-3}
5	*TTHB192*	hypothetical protein	+ 1.669	4.68×10^{-3}
6	*TTHB193*	hypothetical protein	+ 1.446	6.84×10^{-3}
7	*TTHB194*	hypothetical protein	+ 1.549	1.71×10^{-2}

(Operon) Numbers indicate positions of the genes within the operon. (S) Single transcriptional unit. (Role) Biological function. (LogFC) Log2-fold change between data obtained from TTHB099-deficient and wild-type *T. thermophilus* HB8 strains. (Adj. P-value) The P-value obtained following multiple testing corrections using the default Benjamini and Hochberg false discovery rate method [14].

DISCUSSION

In this study, an *in vitro* iterative selection method, REPSA, was used to annotate the TTHB099 transcription regulator in *T. thermophilus* HB8. This, coupled with next generation sequencing and MEME motif elicitation allowed for the identification of the TTHB099-DNA binding motif, a 16 bp long palindromic sequence, 5'–TGT(A/g)n(t/c)c(t/c)(a/g)g(a/g)n(T/c)ACA–3', with a consensus half-site 5'–$T_1G_2T_3(A/G)_4N_5(T/C)_6C_7(T/C)_8$–3'. Binding kinetics between TTHB099 and its consensus sequence, as well as single point mutations within its half-site, were investigated using BLI. TTHB099 protein bound the 16-mer consensus sequence with a high affinity (K_D = 2.21 nM) and the point-mutated sequences in the range of 4.86 of 33.6 nM with mutations at the second and third positions having the greatest effect. The different binding affinities for each mutated sequence mirrored the MEME results represented by the TTHB099 sequence logo. Our report is the first time a consensus sequence has been identified for TTHB099.

Interestingly, our sequence has a strong resemblance to the CRP_{Ec} consensus sequence, 5'-AAATGTGATCTAGATCACATTT-3' [16]. In both cases, the trimers "TGT" and "ACA" are highly conserved and are considered most significant for TF binding. The specifics of this resemblance could be correlated to the homology between the two proteins previously reported by Agari *et al.* [50]. However, *E. coli* and *T. thermophilus* HB8 are not only phylogenetically distant, but they also live in entirely different environments, mesophilic and extremophilic, respectively [69]. Hence, the

biological roles of TTHB099 need not necessarily be the same as those of CRP_{Ec}. This is most evident in the observation that TTHB099 does not require the second messenger 3',5' cAMP to bind DNA, one required by CRP_{Ec}. Considering that *T. thermophilus* HB8 phylogenetic positioning is within the deepest branches close to the last universal common ancestor (LUCA), slower evolutionary changes could explain the differences between its' CRP proteins and CRP_{Ec} [70].

Having found and validated a consensus TTHB099-binding sequence, mapping it into the genome of *T. thermophilus* HB8 would help identify potential TTHB099-regulated genes. Using FIMO, the MEME derived position weight matrix version of our consensus sequence recognized 78 sequences. The top 25 sequences with the best p-values were selected for further validation. It is important to note that the p-values derived were not as small as found in our previous studies due to the ten poorly conserved positions in the middle of the TTHB099 consensus sequence palindrome, which affected the dynamic programming algorithm of FIMO. Our analysis of the TTHB099 binding sites' location relative to the TSS of the proximal downstream genes showed that almost half of the identified sites were located inside open reading frames, which is not typical for traditional transcription factors. Notably, no potential TTHB099 binding site was found near its gene. This could imply that the TTHB099 TF has no direct regulatory role over its operon litR (*TTHB100, TTHB099, TTHB098*) or the divergent crtB operon (*TTHB101, TTHB102*) that shares a common intergenic region. Autoregulation is a common feature for many prokaryotic TFs, including members of the CRP family, but may not be a characteristic for TTHB099 unless in an auxiliary fashion [71].

The promoter analysis revealed that nine TTHB099-binding sites overlapped with potential core promoter elements, a TF-promoter interaction characteristic of Class II transcription activators, as well as transcription inhibition via steric hindrance. Additionally, three sequences bound upstream of the –35 box, fitting the Class I activator model, while two bound downstream of the –10 box, a model used by both transcription activators and repressors. These variations in the binding method suggest that TTHB099 could be either an activator or a suppressor. Indeed, the dual regulatory role is common in global regulators such as CRP_{Ec} [72]. Moreover, eight pairs of the TTHB099-binding sequences were found in the intergenic region of divergent genes, another characteristic of dual-regulators [31].

Biophysical studies performed with BLI were used to further our understanding of TTHB099 interaction with the identified sites. The equilibrium dissociation constants were below the micromolar range, showing that TTHB099 had some appreciable affinity for the tested sites. However, variations as high as 200-fold were observed. These K_D changes did not follow any particular trends, such as the P-value order established through FIMO, neither did the sites with the highest affinity have similarities in terms of promoter location or presumed manner of transcription regulation. For example, the TTHB099 binding sequence with the highest affinity (3.05 nM) was located in the intergenic region and overlapped with the –35 box upstream of *TTHA1833*. The TTHB099 binding sequences with the second-lowest K_D were also situated in the intergenic regions, but they were located upstream and downstream of the *TTHB088/89* promoters, respectively. Such biophysical results emphasize the importance of experimental validation of theoretically determined sites.

Our BLI binding studies are limited to the simple interactions of a purified protein with synthesized DNAs in the absence of any environmental or biological factors. Knowing that the transcription regulation apparatus can be complex, we decided to complement our *in vitro* study with data from *in vivo* expression profiles. Using publicly available expression profile data from the matched wild type and *TTHB099*-deficient *T. thermophilus* HB8 strains, operons of the 16 potentially regulated genes were investigated. We found that the mRNAs of these genes were not significantly affected by the deficiency of TTHB099. Moreover, the biological roles of half of these genes were hypothetical due to the limited studies on gene annotation in the organism (Figure 11). These results suggest that TTHB099 does not have any appreciable regulatory roles over these genes in exponentially propagating wild type organisms.

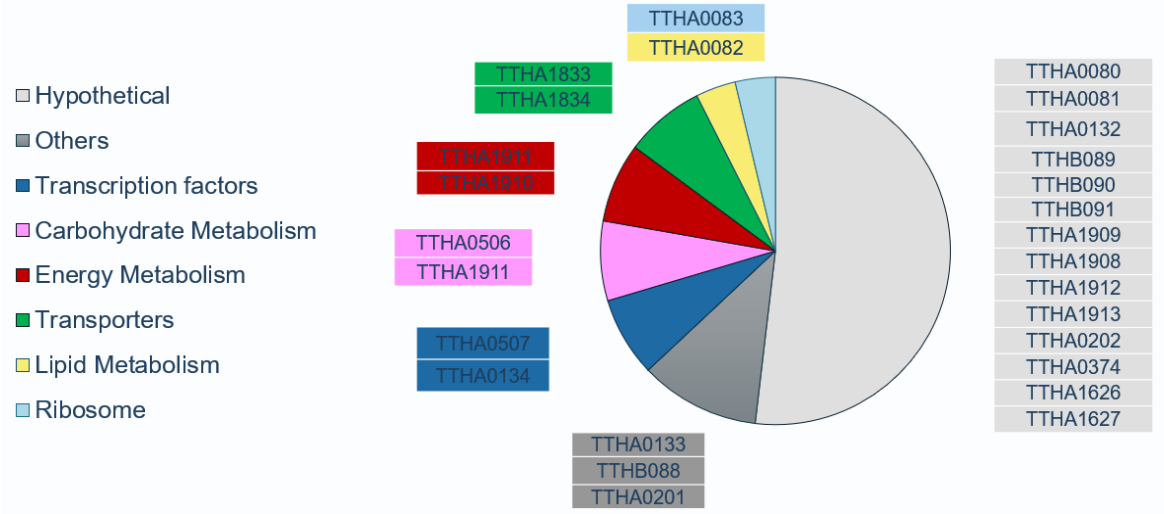

Figure 11. The expression profile for potential TTHB099-regulated genes. Shown is a pie chart of FIMO identified genes and their operons containing TTHB099-binding motif near their regulatory elements organized based on their role in metabolic pathways.

Nonetheless, TTHB099 deficiency does appreciably affect the expression of several genes in exponentially propagating *T. thermophilus* HB8. We identified 19 operons, 12 of

which were overexpressed (positively affected) in the deficient strains. The upregulated set of genes were involved in the electron transport chain (ETC) of oxidative phosphorylation, sugar metabolism, type IV pilin related proteins, and one osmotically inducible protein. These genes were grouped based on their role in various metabolic pathways shown in Figure 12. Most were part of carbohydrate and energy metabolism, followed by signaling, bacterial secretion, and cofactor and vitamin metabolism. A few genes were hypothetical and two were identified as transcription factors, which adds to the complexity of TTHB099 TF's regulatory mechanism.

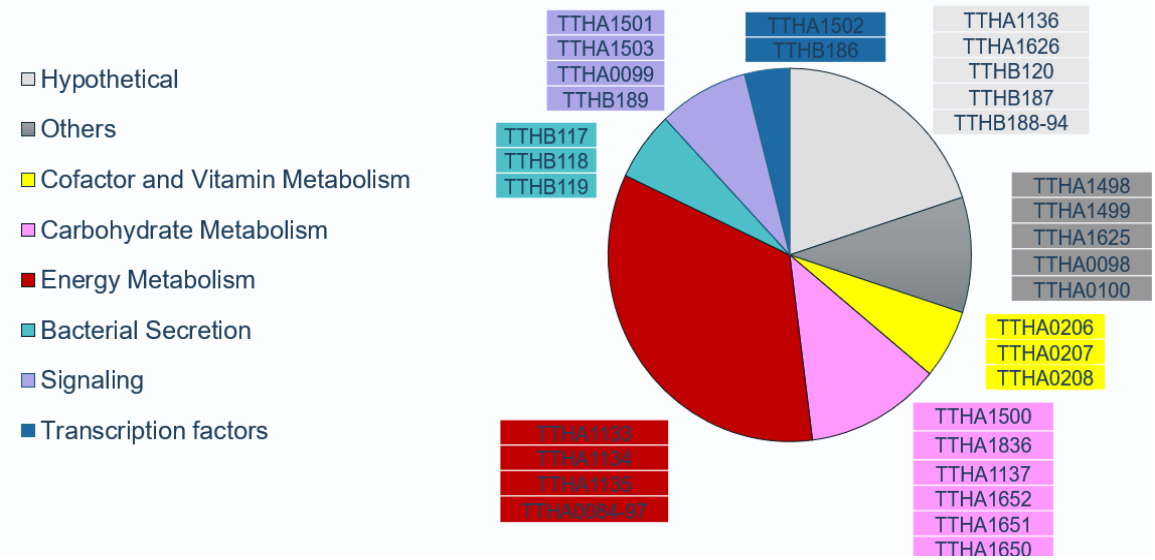

Figure 12. Upregulated operons in *TTHB099*-deficient strain. Shown is a pie chart of the upregulated genes and their operons grouped by their metabolic role.

Conversely, there were seven underexpressed operons or a total of 17 genes in the TTHB099-deficient strains, suggesting that TTHB099 may act as an activator for these genes. The downregulated genes encode for ribosomal proteins, iron ABC transporters, and ATPases. The downregulated operons were grouped in the following metabolic pathways: ribosome, transport, and others (Figure 13). Most genes in the "others" group were singularly involved in protein translation and post-translational modifications.

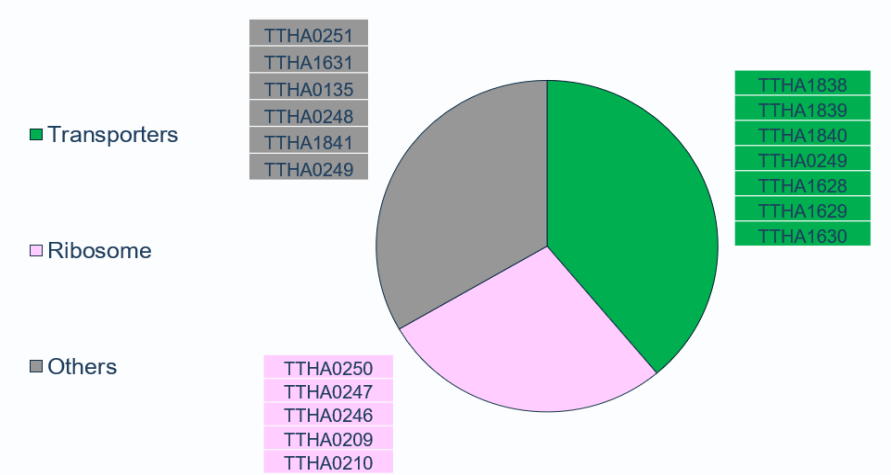

Figure 13. Downregulated operons in *TTHB099*-deficient strain. Shown is a pie chart of downregulated genes and their operons grouped by their metabolic pathways.

Notably, the biological roles of the most affected operons in the *TTHB099*-deficient strain were involved in metabolic pathways that have been reported to be regulated by CRP_{Ec} [73]. For example, ribosome-related genes were downregulated in the *TTHB099*-deficient strain, similar to what Pal *et al*. reported for their evolutionary expressed CRP_{Ec}-deficient strains [74]. Likewise, iron transport genes were downregulated in the *TTHB099*-depleted strain, similar to what was observed in the absence of CRP_{Ec}, as Zhang et al. reported [75]. Such results indicate that TTHB099 does have some biological functions like those of the CRP_{Ec}. However, these regulatory roles do not seem to be affected by changes in cAMP concentration. Moreover, a MEME search for a consensus sequence between the 19 most-affected operons identified via the GEO data failed to bring up any significant motifs. Thus, the hypothesis for a simple regulatory mechanism is once more unsatisfied.

TT_P0055 from *T. thermophilus* HB27, an ortholog of TTHB099 with only one aa substitution (E77D), has been reported to be a positive regulator of *crtB* operon, which in turn is involved in light-dependent carotenoid biosynthesis [48]. However, the functional

effects of TT_P0055 on carotenoid production lack details on the mechanism of regulation and could indicate that TT_P0055 has indirect control over *crtB* activation. The homology between the HB27 and HB8 strains, particularly on this regulatory complex (TT_P0055 and TTHB099 proteins, their intergenic regions, and their *crtB* operons), would suggest similar biological functions for the two TFs. However, when analyzing the GEO expression data in the *TTHB099*-deficient strain, there is no detectable change in *crtB* genes. These results could be attributed to the absence of light in the experimental conditions required to deplete the litR transcriptional repressor of TT_P0055, the latter positively regulating carotenoid production [48]. Further profile expression data under different environmental conditions are necessary to correlate phenotypic results with those from the mRNA expressions.

Because TTHB099 does not seem to have any observable binding to the *PcrtB* promoter, the study published by Ebright *et al*. centered on TTHB099 binding upstream of *TTHB101* is based on a prediction not firmly established [76]. Hence, Ebright's claim that TTHB099 is a model class II transcription activator may need to be reconsidered under the light of our new findings.

Looking for a connection between the genes found via the REPSA-identified consensus sequence and the genes affected by TTHB099 deficiency, as determined by GEO2R, we found that five of the affected operons (30 genes) had an upstream binding sequence identified by FIMO. These binding sites were located at about 0.9 to 4 kbp away upstream of the most affected operons. Such behavior could be explained by TTHB099 acting as an enhancer or silencer. These elements do exist in the prokaryotic world but not in large numbers. To date, the identified prokaryotic enhancers regulate

only a few promoters used by σ^{54} -directed RNA polymerases [77]. Knowing that *T. thermophilus* HB8 does not have a σ^{54} homolog, it becomes even more challenging to predict the mechanism of action for TTHB099 as an enhancer/silencer. Future studies could be designed to analyze potential interactions of TTHB099 with other TFs, supporting the hypothesis of a complex regulatory mechanism involving distal enhancer/silencer elements. As for TTHB099 being an activator or a suppressor, all our data point towards a dual regulatory role.

CONCLUSION

Our reverse genetic approach determined the preferred DNA-binding sequence for TTHB099 TF, the 16 bp long palindromic sequence 5'–TGT(A/g)n(t/c)c(t/c)(a/g)g(a/g)n(T/c)ACA–3'. These findings encouraged the mapping of this sequence into the genome of *T. thermophilus* HB8, where 25 potential target genes were identified. Binding kinetics studies coupled with bioinformatics studies of transcription regulators' common attributes led to 16 ideal targets. We complemented our analysis with publicly available *in vivo* expression data. We observed that our 16 target genes and respective operons were not significantly up or downregulated in the *TTHB099* deficient mutant. However, 19 operons without any identified consensus sequence were affected in the mutated strain. We predict a complex regulatory mechanism for TTHB099 in *T. thermophilus* HB8, most probably in a dual-regulator role. This study has been published in the *International Journal of Molecular Sciences*, and the supplemental material can be found at https://doi.org/10.3390/ijms21217929.

Future studies could include more expression profile experiments under different environmental conditions, starting with the effects of light on *T. thermophilus* HB8. Following the study on the closely related *T. thermophilus* HB27, where the organism experienced phenotypical variations in light and dark conditions, it would be interesting to see if HB8 will display similar changes. Moreover, would the mutation of the same gene or operon in both strains lead to similar or different phenotypic responses?

The positioning of *T. thermophilus* HB8 close to LUCA in the phylogenetic tree could suggest that this organism shares similarities with archaea, specifically thermophilic archaea. Further bioinformatic studies, in particular genome comparison studies, could reveal more about the evolution of *T. thermophilus* HB8 organism and TTHB099's role in transcription regulation. For instance, since TTHB099 does not require cAMP to bind DNA, what other factors allow this TF to regulate transcription according to environmental changes? Could it be changes in TTHB099 concentration influencing promoter regions with various affinities? Moreover, are these factors similar to what bacteria use or more like what archaea employ? The following answers would complement this study, as well as provide a better understanding of the regulatory mechanisms for TTHB099 and other CRP like proteins in prokaryotic organism.

REFERENCES

1. Seshasayee, A. S. N.; Sivaraman, K.; Luscombe, N. M. In An overview of prokaryotic transcription factors; A Handbook of Transcription Factors. *Springer* **2011**,7-23.

2. Browning, D. F.; Busby, S. J. The regulation of bacterial transcription initiation. *Nat. Rev. Microbiol.* **2004**, *2*, 57-65.

3. Estrem, S. T.; Gaal, T.; Ross, W.; Gourse, R. L. Identification of an UP element consensus sequence for bacterial promoters. *Proc. Natl. Acad. Sci. U.S.A.* **1998***, 95*, 9761-9766.

4. Hook-Barnard, I. G.; Hinton, D. M. Transcription initiation by mix and match elements: flexibility for polymerase binding to bacterial promoters. *Gene Regul. Syst. Biol.* **2007***, 1*, 117762500700100020.

5. Dorman, C. J.; Co-operative roles for DNA supercoiling and nucleoid-associated proteins in the regulation of bacterial transcription**.** *Biochem. Soc. Trans.* **2013***, 41*, 542–547.

6. Borukhov, S.; Nudler, E. RNA polymerase holoenzyme: structure, function and biological implications. *Curr. Opin. Microbiol.* **2003***, 6*, 93-100.

7. Collado-Vides, J.; Magasanik, B.; Gralla, J. D. Control site location and transcriptional regulation in Escherichia coli. *Microbiol. Mol. Biol. Rev.* **1991***, 55*, 371-394.

8. Mejía-Almonte, C.; Busby, S. J.; Wade, J. T.; van Helden, J.; Arkin, A. P.; Stormo, G. D.; Eilbeck, K.; Palsson, B. O.; Galagan, J. E.; Collado-Vides, J. Redefining fundamental concepts of transcription initiation in bacteria. *Nat. Rev. Genet.* **2020**, 1-16.

9. Xu, H.; Hoover, T. R. Transcriptional regulation at a distance in bacteria. *Curr. Opin. Microbiol.* **2001***, 4*, 138-144.

10. Browning, D. F.; Butala, M.; Busby, S. J. Bacterial Transcription Factors: Regulation by Pick "N" Mix. *J. Mol. Biol.* **2019**.

11. Perez-Rueda, E.; Hernandez-Guerrero, R.; Martinez-Nuñez, M. A.; Armenta-Medina, D.; Sanchez, I.; Ibarra, J. A. Abundance, diversity, and domain architecture variability in prokaryotic DNA-binding transcription factors. *PLoS ONE* **2018**, 13.

12. Sahota, G.; Stormo, G. D. Novel sequence-based method for identifying transcription factor binding sites in prokaryotic genomes. *Bioinformatics* **2010***, 26*, 2672-2677.

13. Pérez-Rueda, E.; Collado-Vides, J.; Segovia, L. Phylogenetic distribution of DNA-binding transcription factors in bacteria and archaea. *Comput. Biol. Chem.* **2004**, *28*, 341-350.

14. Schreiter, E. R.; Drennan, C. L. Ribbon–helix–helix transcription factors: variations on a theme. *Nat. Rev. Microbiol.* **2007***, 5*, 710-720.

15. Charoensawan, V.; Wilson, D.; Teichmann, S. A. Genomic repertoires of DNA-binding transcription factors across the tree of life. *Nucleic Acids Res.* **2010***, 38*, 7364-7377.

16. Leuze, M. R.; Karpinets, T. V.; Syed, M. H.; Beliaev, A. S.; Uberbacher, E. C. Binding motifs in bacterial gene promoters modulate transcriptional effects of global regulators CRP and ArcA. *Gene Regul. Syst. Biol.* **2012***, 6*, S9357.

17. Martin, R. G.; Gillette, W. K.; Rhee, S.; Rosner, J. L. Structural requirements for marbox function in transcriptional activation of mar/sox/rob regulon promoters in Escherichia coli:

sequence, orientation and spatial relationship to the core promoter. *Mol. Microbiol.* **1999***, 34*, 431-441.

18. Zenkin, N.; Yuzenkova, Y. New insights into the functions of transcription factors that bind the RNA polymerase secondary channel. *Biomolecules* **2015***, 5*, 1195-1209.
19. Dian, C. Adaptive Responses by Transcriptional Regulators to small molecules in Prokaryotes: Structural studies of two bacterial one-component signal transduction systems DntR and HpNikR. *Stockholm University,* **2007**.
20. Gomelsky, M. cAMP, c-di-GMP, c-di-AMP and now cGMP: bacteria use them all! *Mol. Microbiol.* **2011***, 79*, 562-565.
21. Mironov, A. S.; Gusarov, I.; Rafikov, R.; Lopez, L. E.; Shatalin, K.; Kreneva, R. A.; Perumov, D. A.; Nudler, E. Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria. *Cell* **2002***, 111*, 747-756.
22. Rodionov, D. A.; Dubchak, I. L.; Arkin, A. P.; Alm, E. J.; Gelfand, M. S. Dissimilatory metabolism of nitrogen oxides in bacteria: comparative reconstruction of transcriptional networks. *PLoS Comput Biol.* **2005***, 1*.
23. Nguyen, C. C.; Saier, M. H. Phylogenetic, structural and functional analyses of the LacI-GalR family of bacterial transcription factors. *FEBS Lett.* **1995***, 377*, 98-102.
24. Schujman, G. E.; Paoletti, L.; Grossman, A. D.; de Mendoza, D. FapR, a bacterial transcription factor involved in global regulation of membrane lipid biosynthesis. *Dev. Cell* **2003***, 4*, 663-672.
25. Martínez-Antonio, A.; Collado-Vides, J. Identifying global regulators in transcriptional regulatory networks in bacteria. *Curr. Opin. Microbiol*. **2003**, *6*, 482-489.
26. Moreno-Campuzano, S.; Janga, S. C.; Pérez-Rueda, E. Identification and analysis of DNA-binding transcription factors in Bacillus subtilis and other Firmicutes-a genomic approach. *BMC Genomics* **2006***, 7*, 147.
27. Rodionov, D. A. Comparative genomic reconstruction of transcriptional regulatory networks in bacteria. *Chem. Rev.* **2007***, 107*, 3467-3497.
28. van Hijum, S. A.; Medema, M. H.; Kuipers, O. P. Mechanisms and evolution of control logic in prokaryotic transcriptional regulation. *Microbiol. Mol. Biol. Rev.* **2009***, 73*, 481-509.
29. Munson, G. P.; Scott, J. R. Rns, a virulence regulator within the AraC family, requires binding sites upstream and downstream of its own promoter to function as an activator. *Mol. Microbiol.* **2000***, 36*, 1391-1402.
30. Ishihama, A. Prokaryotic genome regulation: a revolutionary paradigm. *Proc. Japan Acad. , Ser. B.* **2012***, 88*, 485-508.
31. Balleza, E.; Lopez-Bojorquez, L. N.; Martínez-Antonio, A.; Resendis-Antonio, O.; Lozada-Chávez, I.; Balderas-Martínez, Y. I.; Encarnación, S.; Collado-Vides, J. Regulation by transcription factors in bacteria: beyond description. *FEMS Microbiol. Rev.* **2008***, 33*, 133-151.
32. Cases, I.; De Lorenzo, V.; Ouzounis, C. A. Transcription regulation and environmental adaptation in bacteria. *Trends Microbiol.* **2003***, 11*, 248-253.

33. Ali, F.; Seshasayee, A. S. N. Dynamics of genetic variation in transcription factors and its implications for the evolution of regulatory networks in Bacteria. *Nucleic Acids Res.* **2020***, 48*, 4100-4114.

34. Ma, C.; Yang, X.; Lewis, P. J. Bacterial transcription as a target for antibacterial drug development. *Microbiol. Mol. Biol. Rev.* **2016***, 80*, 139-160.

35. Inukai, S.; Kock, K. H.; Bulyk, M. L. Transcription factor–DNA binding: beyond binding site motifs. *Curr. Opin. Genet. Dev.* **2017***, 43*, 110-119.

36. Chen, D.; Orenstein, Y.; Golodnitsky, R.; Pellach, M.; Avrahami, D.; Wachtel, C.; Ovadia-Shochat, A.; Shir-Shapira, H.; Kedmi, A.; Juven-Gershon, T. SELMAP-SELEX affinity landscape MAPping of transcription factor binding sites using integrated microfluidics. *Sci. Rep.* **2016***, 6*, 33351.

37. Zhuo, Z.; Yu, Y.; Wang, M.; Li, J.; Zhang, Z.; Liu, J.; Wu, X.; Lu, A.; Zhang, G.; Zhang, B. Recent advances in SELEX technology and aptamer applications in biomedicine. *Int. J. Mol. Sci.* **2017***, 18*, 2142.

38. Stormo, G. D.; Zhao, Y. Determining the specificity of protein–DNA interactions. *Nat. Rev. Genet.* **2010***, 11*, 751-760.

39. Van Dyke, M. W.; Van Dyke, N.; Sunavala-Dossabhoy, G. REPSA: General combinatorial approach for identifying preferred ligand–DNA binding sequences. *Methods* **2007***, 42*, 118-127.

40. Van Dyke, M. W.; Beyer, M. D.; Clay, E.; Hiam, K. J.; McMurry, J. L.; Xie, Y. Identification of preferred DNA-binding sites for the Thermus thermophilus transcriptional regulator SbtR by the combinatorial approach REPSA. *PLoS ONE.* **2016***, 11*, e0159408.

41. Van Dyke, M. Design and PCR synthesis of infrared fluorophore-labeled modular DNA probes. *Protocols.io* **2018**, doi: 10.17504/protocols.io.wfjfbkn

42. Oshima, T.; Imahori, K. Description of Thermus thermophilus (Yoshida and Oshima) comb. nov., a nonsporulating thermophilic bacterium from a Japanese thermal spa. *Int. J. Syst. Evol. Microbiol*. **1974**, *24*, 102-112.

43. Masui, R.; Kurokawa, K.; Nakagawa, N.; Tokunaga, F.; Koyama, Y.; Shibata, T.; Oshima, T.; Yokoyama, S.; Yasunaga, T.; Kuramitsu, S. Complete genome sequence of Thermus thermophilus HB8. *NCBI http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi* **2004**.

44. Yokoyama, S.; Hirota, H.; Kigawa, T.; Yabuki, T.; Shirouzu, M.; Terada, T.; Ito, Y.; Matsuo, Y.; Kuroda, Y.; Nishimura, Y. Structural genomics projects in Japan. *Nat. Struct. Mol. Biol.* **2000***, 7*, 943.

45. Mohamed, N.; Elfaitouri, A.; Fohlman, J.; Friman, G.; Blomberg, J. A sensitive and quantitative single-tube real-time reverse transcriptase-PCR for detection of enteroviral RNA. *J. Clin. Virol.* **2004**, *30*, 150-156.

46. Myers, T. W.; Gelfand, D. H. Reverse transcription and DNA amplification by a Thermus thermophilus DNA polymerase. *Biochem.* **1991**, *30*, 7661-7666.

47. Kanehisa, M.; Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. **2000**, *28*, 27-30.

48. Takano, H.; Kondo, M.; Usui, N.; Usui, T.; Ohzeki, H.; Yamazaki, R.; Washioka, M.; Nakamura, A.; Hoshino, T.; Hakamata, W. Involvement of CarA/LitR and CRP/FNR family transcriptional regulators in light-induced carotenoid production in Thermus thermophilus. *J. Bacteriol.* **2011***, 193*, 2451-2459.

49. UniProt Consortium UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **2018***, 47*, D506-D515.

50. Agari, Y.; Kuramitsu, S.; Shinkai, A. X-ray crystal structure of TTHB099, a CRP/FNR superfamily transcriptional regulator from Thermus thermophilus HB8, reveals a DNA-binding protein with no required allosteric effector molecule. *Proteins.* **2012***, 80*, 1490-1494.

51. Parkinson, G.; Wilson, C.; Gunasekera, A.; Ebright, Y. W.; Ebright, R. E.; Berman, H. M. Structure of the CAP-DNA complex at 2.5 Å resolution: a complete picture of the protein-DNA interface. *J. Mol. Biol.* **1996***, 260*, 395-408.

52. Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004***, 25*, 1605-1612.

53. Brüggemann, H.; Chen, C. Comparative genomics of Thermus thermophilus: plasticity of the megaplasmid and its contribution to a thermophilic lifestyle. *J. Biotechnol.* **2006***, 124*, 654-661.

54. Ramírez-Arcos, S.; Fernández-Herrero, L. A.; Marín, I.; Berenguer, J. Anaerobic growth, a property horizontally transferred by an Hfr-like mechanism among extreme thermophiles. *J. Bacteriol.* **1998***, 180*, 3137-3143.

55. Van Dyke, M. Direct double-stranded DNA quantitation from PCR reactions. *Protocols.io* **2017**, doi: 10.17504/protocols.io.k5pcy5n

56. Van Dyke, M.; Gracien, I. Restriction endonuclease protection assays using infrared-fluorescent probes. *Protocols.io* **2020**, bi5ikg4e, doi:10.17504/protocols.io.bi5ikg4e.

57. Bailey, T. L.; Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in bipolymers. **1994**.

58. Van Dyke, M.; Cox, J. Electrophoretic mobility shift assays using infrared-fluorescent DNA probes. *Protocols.io*, **2018**, doi: 10.17504/protocols.io.mbdc2i6

59. Grant, C. E.; Bailey, T. L.; Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **2011**, *27*, 1017-1018.

60. Salamov, V. S. A.; Solovyevand, A. Automatic annotation of microbial genomes and metagenomic sequences. Metagenomics and its applications in agriculture, biomedicine and environmental studies. *Hauppauge: Nova Science Pub.* **2011**, 61-78.

61. Taboada, B.; Ciria, R.; Martinez-Guerrero, C.E.; Merino, E. ProOpDB: Prokaryotic Operon DataBase. *Nucleic Acids Res.* **2012**, *40*, D627–D631.

62. Karp, P. D.; Billington, R.; Caspi, R.; Fulcher, C. A.; Latendresse, M.; Kothari, A.; Keseler, I. M.; Krummenacker, M.; Midford, P. E.; Ong, Q. The BioCyc collection of microbial genomes and metabolic pathways. *Brief. Bioinformatics* **2019**, *20*, 1085-1093.

63. Barrett, T.; Wilhite, S.E.; Ledoux, P.; Evangelista, C.; Kim, I.F.; Tomashevsky, M.; Marshall, K.A.; Phillippy, K.H.; Sherman, P.M.; Holko, M.; et al. NCBI GEO: Archive for functional genomics data sets—update. *Nucleic Acids Res.* **2013**, *41*, D991–D995.

64. Lee, M.; Um, H.; Van Dyke, M. W. Identification and characterization of preferred DNA-binding sites for the Thermus thermophilus transcriptional regulator FadR. *PLoS ONE.* **2017***, 12*, e0184796.

65. Cox, J. S.; Moncja, K.; Mckinnes, M.; Van Dyke, M. W. Identification and Characterization of Preferred DNA-Binding Sites for the Thermus thermophilus HB8 Transcriptional Regulator TTHA0973. *Int. J. Mol. Sci.* **2019***, 20*, 3336.

66. Cox, J. S.; Van Dyke, M. W. General and Genomic DNA-Binding Specificity for the Thermus thermophilus HB8 Transcription Factor TTHB023. *Biomolecules* **2020***, 10*, 94.

67. Popovych, N.; Tzeng, S.R.; Tonelli, M.; Ebright, R.H.; Kalodimos, C.G. Structural basis for cAMP-mediated allosteric control of the catabolite activator protein. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 6927–6932.

68. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B* **1995**, *57*, 289–300.

69. Hartmann, R. K.; Wolters, J.; Kröger, B.; Schultze, S.; Specht, T.; Erdmann, V. A. Does Thermus represent another deep eubacterial branching? *Syst. Appl. Microbiol.* **1989***, 11*, 243-249.

70. Hartmann, R. K.; Wolters, J.; Kröger, B.; Schultze, S.; Specht, T.; Erdmann, V. A. Does Thermus represent another deep eubacterial branching? *Syst. Appl. Microbiol.* **1989***, 11*, 243-249.

71. Aiba, H. Autoregulation of the Escherichia coli crp gene: CRP is a transcriptional repressor for its own gene. *Cell* **1983***, 32*, 141-149.

72. Galan-Vasquez, E.; Sanchez-Osorio, I.; Martinez-Antonio, A. Transcription factors exhibit differential conservation in bacteria with reduced genomes. *PLoS ONE* **2016***, 11*, e0146901.

73. Shimada, T.; Fujita, N.; Yamamoto, K.; Ishihama, A. Novel roles of cAMP receptor protein (CRP) in regulation of transport and metabolism of carbon sources. *PLoS ONE* **2011***, 6*, e20081.

74. Pal, A.; Iyer, M.S.; Srinivasan, S.; Seshasayee, A.S.; Venkatesh, K.V. Elucidating the regulatory role of CRP in coordinating protein biosynthesis machinery with metabolism that defines growth optimality in Escherichia coli. *bioRxiv* **2020**, doi:10.1101/2020.06.18.159491.

75. Zhang, Z.; Gosset, G.; Barabote, R.; Gonzalez, C. S.; Cuevas, W. A.; Saier, M. H. Functional interactions between the carbon and iron utilization regulators, Crp and Fur, in Escherichia coli. *J. Bacteriol.* **2005***, 187*, 980-990.

76. Feng, Y.; Zhang, Y.; Ebright, R. H. Structural basis of transcription activation. *Science* **2016***, 352*, 1330-1333.

77. Bondarenko, V.; Liu, Y.; Ninfa, A.; Studitsky, V. M. Action of prokaryotic enhancer over a distance does not require continued presence of promoter-bound σ54 subunit. *Nucleic Acids Res.* **2002***, 30*, 636-642.